Формальная модель русской морфологии (словоизменения)

Словарь Зализняка

Одним из широкодоступных (и активно используемых) русскоязычных ЛБД является электронный вариант фундаментального «Грамматического словаря русского языка» А.А.Зализняка. Текст словаря был перенесен на машинные носители в начале 80-х годов. С тех пор словари всех русскоязычных текстовых процессоров, словари практически всех экспериментальных и коммерческих систем машинного перевода и других АОТ-систем строятся на основе словаря Зализняка.

Полиграфический вариант словаря Зализняка состоит из двух частей: «Грамматические сведения» (около 120 страниц) и собственно «Словарь» (около 740 страниц). В первой части представлена разработанная автором словаря с необычайной тщательностью оригинальная модель русского словоизменения (склонения и спряжения). Во второй приведено около 100 тысяч слов, которым приписаны грамматические индексы, характеризующие тип их словоизменения и схему ударения. Слова упорядочены по концам, что естественно и удобно для грамматического словаря, поскольку слова со сходным грамматическим поведением (одинаковыми суффиксами и окончаниями) располагаются компактными группами.

Словарная статья в словаре Зализняка состоит из заголовка (начальная форма слова) и словарной (грамматической) информации. Для некоторых слов даются также дополнительные сведения, необходимые для различения вариантов. Статьи с заголовками *лев, стричь* и *прихожая* выглядят так:

```
лев мо 1*b (животное)

лев м 1a (денежная единица)

стричь нсв 8b (-г-)

прихожая ж (п 4a)
```

По первому элементу словарной информации определяется грамматический класс (***спрягаемое*** слово, слово ***субстантивного, адъективного*** или ***местоименного*** склонения эти термины будут разъяснены позже), для слов субстантивного склонения также одушевленность и род, для спрягаемых слов – вид. Если первый элемент – «п», то слово относится к словам адъективного склонения; «ж» – к словам субстантивного склонения, женского рода, неодушевленным; «мо» – к словам субстантивного склонения, мужского рода, одушевленным; «нсв» – к спрягаемым словам (глаголам) несовершенного вида.

Если второй элемент – не цифра, то это означает, что слово изменяется по необычной модели (существительное *прихожая* изменяется по модели слов адъективного склонения). Остальные элементы словарной статьи либо уточняют тип склонения/спряжения, либо свидетельствуют о наличии в слове чередований (символ *), об отсутствии у слова некоторых форм или о других частных особенностях словоизменения. Буквенный индекс после цифры (или после символа *) характеризует схему ударения во всех формах описываемого слова; эта информация полезна при автоматизированной генерации фонетического словаря словоформ русского языка.

Отметим, что исходный (полиграфический) вариант словаря Зализняка был ориентирован на пользователя-человека. Основной сценарий использования словаря предусматривал возможность просклонять/проспрягать любое слово из «Словаря» на основе его грамматического описания и правил, приведенных в «Грамматических сведениях». Эти операции, вообще говоря, требовали выполнения некоторых трудноформализуемых действий, определенной языковой компетенции: поиск уместных грамматических таблиц, определение типа чередования, рассуждения по аналогии. Поэтому непосредственное использование словаря Зализняка (даже в электронном виде) в составе компьютерных систем обработки текста/речи затруднительно.

Разработчики компьютерных словарей, базирующихся на словаре Зализняка, выбирают обычно один из трех путей:

- генерация на основе словаря Зализняка словаря русских словоформ;

- использование электронного «Словаря» в исходной форме и разработка (достаточно сложных) алгоритмов, моделирующих работу с «Грамматическими сведениями»;

- создание на основе словаря Зализняка формальной модели словоизменения и необходимое переструктурирование словарной части (явное введение в словарную статью некоторой информации из «Грамматических сведений»), позволяющее существенно упростить алгоритмы.

После подобных преобразований компьютерный словарь может использоваться для решения двух практически важных задач:

- задача морфологического анализа – определения начальной формы слова по произвольной словоформе (и, возможно, грамматических признаков словоформы);
- задача синтеза – построения всех форм (или указанной формы) слова по начальной форме.

Одна из первых формальных моделей русского словоизменения на базе словаря Зализняка (третий из указанных выше путей) была разработана еще в середине 80-х годов на кафедре алгоритмических языков факультета ВМК МГУ под руководством М.Г.Мальковского. Модель была реализована на лиспоподобном языке программирования Плэнер (ЭВМ БЭСМ-6, а позже – МВК «Эльбрус-2» и IBM-совместимые ПК). При этом широко использовались динамические структуры, мощные средства обработки списков и сопоставления образца с выражением.

В плэнерских структурах данных явно указывались все морфологические свойства для каждого слова, включая чередования в основе слова. Поэтому плэнерское представление достаточно легко воспринималось человеком, явно отражало морфологические особенности описываемых в компьютерном словаре слов.

Однако язык Плэнер является интерпретируемым, а следовательно, довольно медленно работающим, что затрудняет его применение в системах, к которым предъявляются высокие требования по быстродействию. Обработка сложной структуры списков требует существенных затрат машинного времени, даже при реализации алгоритма их обработки на компилируемых языках, ориентированных

на написание эффективных программ (C, C++). Поэтому было принято решение о переходе к другой структуре словаря и соответствующей модификации алгоритмов анализа и синтеза.

Плэнерские структуры, описывающие морфологические особенности всех различных классов слов, были пронумерованы. Затем словам/основам и флексиям были сопоставлены соответствующие номера классов. При чередовании в основе и при наличии у слова супплетивных форм, образованных от другой основы (*хороший – лучше*), были организованы дополнительные входы в словарные статьи.

Новое представление словаря трудно воспринимаемо для человека. Однако унификация и упрощение структур данных позволили создать условия для значительного увеличения скорости обработки.

Формализация русского словоизменения

В ***Формальной модели русского словоизменения*** (ФМРС) множество слов русского языка разбивается на два основных класса – ***неизменяемые*** (Н-слова) и ***изменяемые***, т.е. склоняемые или спрягаемые (И-слова). Совокупность форм И-слова (словоформ) образует его ***парадигму***. В каждой словоформе можно выделить ***основу*** и окончание, или ***флексию*** (возможно, пустую, которую мы обозначим: -∅), соответствующую конкретной форме И-слова; за флексией может следовать ***постфикс***, например, возвратная частица *ся*/*сь*.

С основой И-слова, Н-словом, флексией и словоформой связывается описание значения соответствующего объекта, включающее описание его грамматических характеристик; лексических связей (синонимы, производные слова); семантического значения (ассоциированные с объектом понятия). Грамматические характеристики определяют сочетаемость основ и флексий и синтаксические признаки объектов всех четырех типов.

К грамматическим характеристикам морфологического уровня относятся:

морфологический *(словоизменительный)* ***класс*** – **М-класс**, ***парадигматический класс*** – **П-класс**, ***чередование***, ***исключение***. Синтаксическим показателем является ***синтаксический класс*** (**С-класс**). Если М-класс определяет, как изменяется слово (склоняется, спрягается), то С-класс характеризует его синтаксическое поведение (сочетаемость с другими словами) Как словоизменительные, так и синтаксические признаки определяются набором значений грамматических переменных.

Грамматическая переменная (ГП) – переменная одного из следующих типов: одушевленность, род, число, падеж, вид, лицо, залог, возвратность, время, наклонение, степень – принимает закодированное целым числом значение из некоторого множества допустимых. Значение ГП **род**, например, кодируется так: мужской – 1, женский – 2, средний – 3. Если значение неопределенно, указывается список возможных значений или число 0 (обозначает любое допустимое значение ГП).

Совокупность ГП, по которым изменяется И-слово (***свободных ГП***), определяет его парадигму, а спектр значений этих переменных – число элементов парадигмы. Множество И-слов с общим набором ГП, общим набором свободных ГП и общим спектром значений переменных образует М-класс. Основе (и словоформе)

сопоставлен упорядоченный набор (вектор) значений соответствующих ГП. Так, например, с основой *лев-* слова *лев* (денежная единица) связан такой вектор **(7 8 2 1 0 0)** – слово 7-го М-класса, 8-го П-класса, неодушевленное (2), мужского рода (1), значения ГП **число** и **падеж** не определены (0 и 0). Для словоформы *левами* вектор будет иметь вид **(7 2 1 2 5)**, здесь добавились значения ГП **число** (**2** – множественное) и **падеж** (**5** – творительный).

Понятие М-класса является уточнением традиционного понятия «часть речи»: 7-й класс образован в основном существительными, 8-й − прилагательными, 9-й − глаголами. В ФМРС рассматриваются три класса склоняемых И-слов: местоименные (М-класс номер 5), субстантивные (класс 7), адъективные (класс 8) и один класс спрягаемых (класс 9). Представители 5-го и 8-го М-классов изменяются по родам, числам и падежам, 7-го − по числам и падежам, 9-го − по лицам, родам, числам и временам. Отсутствие у И-слова одной или нескольких форм (например, формы родит. падежа множ. числа у слова *мгла*, форм ед. числа у слова *ножницы*) не мешает отнести его к М-классу номер 7.

Подмножество М-класса, представители которого при совпадающих значениях свободных ГП имеют одинаковые флексии, образует парадигматический класс. В ФМРС рассматриваются 24 П-класса для слов субстантивного склонения, 8 − для слов адъективного склонения, 2 − для слов местоименного склонения, 9 − для спрягаемых слов. К 1-му П-классу субстантивных И-слов относятся, например, существительные *завод* и *артист* (флексии: -∅, *-а*, *-у*, -∅ или *-а*, *-ом*, *-е* − для шести традиционных падежей единственного числа; *-ы*, *-ов*, *-ам*, *-ы* или *-ов*, *-ами*, *-ах* − для множественного); к 11-му П-классу − *карта* и *корова*; к 21-му − *болото*. К 1-му П-классу местоименных И-слов относятся: притяжательное прилагательное *отцов*, существительное *кабельтов* (не изменяется по родам), ко 2-му П-классу − местоимение *мой*, прилагательное *лисий*, порядковое числительное *третий*.

Хотя П-классы задают более детальную классификацию сочетаемости основ с флексиями чем традиционные типы склонения и спряжения, они недостаточны для описания многих частных особенностей русского словоизменения. Эти особенности можно было бы учесть с помощью еще более дробной классификации П-классов, однако в ФМРС используются другие методы.

Как исключения описываются случаи сочетания основы с «нестандартной» для данного П-класса и данной формы флексией: *-а* в форме именит. падежа множеств. числа существительных вместо характерной для 1-го П-класса флексии *-ы* (*глаза*, но *заводы*), пустая флексия вместо флексии *-ов* в родит. падеже множеств. числа (*глаз*, но *заводов*). Исключением считается и наличие у некоторых существительных 2-го родительного (партитивного) и 2-го предложного (локативного) падежей: *кусок сахару*, *в шкафу*, но *из сахара*, *о шкафе*. Всего в ФМРС учитываются 26 исключений такого вида.

К особенностям словоизменения относятся и чередования в основе. В ФМРС учтено 55 чередований, например: *ова - у* (*рис-**ова**-ть - рис-**у**-ю*), *та - щ* (*клеве-**та**-ть - клеве-**щ**-у*), *е* - <пусто> (*царев-**е**-н - царев-н-а*). Для И-слов с чередованиями достаточно рассматривать только один «стандартный» вариант основы, указывая тип и контекст чередования в описании значения основы. Так, для стандартного варианта основы *царевн-* указывается, что при пустой флексии перед последней буквой основы вставляется буква *е*. Относительно редкие чередования

(встречающиеся у 1-3 слов) в ФМРС учитываются по-иному: парадигмы таких слов задаются несколькими основами и Н-словами, образующими ***семейство*** слова (основы *зай-*, *зайд-* и *заш-* и Н-слово *зайти* для глагола *зайти*). Семейства вводятся и для слов с супплетивными формами (*хороший - лучше*) или уникальными наборами флексий (некоторые числительные, личные местоимения).

В синтаксический класс объединяются слова и конструкции с общим набором ГП и общими синтаксическими функциями. Каждому представителю некоторого С-класса сопоставлен (как и в случае М-классов) вектор значений характерных ГП. Для большинства И-слов номер С-класса и набор ГП совпадают с номером и набором ГП М-класса. Так, многие существительные – С-класс номер 7 – относятся и к 7-му М-классу. Однако некоторые слова изменяются по «необычной» модели: существительные *прохожий*, *гончая* склоняются как представители 8-го М-класса.

Библиотека программ «РУССКАЯ МОРФОЛОГИЯ»

Основные программы

Морфологический анализ знакомых слов. Программа МОРФ1

Программа МОРФ1 строит все возможные разбиения входной словоформы на основу и флексию и ищет соответствующие части в словаре (первоначально МОРФ1 пытается найти в словаре совпадающее со словоформой Н-слово, а затем последовательно рассматривает словоформу как основу с пустой флексией, основу с флексиями длиной 3, 2 и 1) или неизменяемое слово.

Проверку правильности разбиения – сочетаемости основы и флексии – осуществляет вспомогательная программа, она же устанавливает значения ГП, определяемые флексией. Когда МОРФ1, отщепив флексию, не может найти полученную основу в словаре, происходит обращение к подпрограмме, применяющей к основе правила чередования. Если и после применения правил чередования найти основу в словаре не удалось, слово признается незнакомым и формируется обращение к программе морфологического анализа незнакомых слов МОРФ2 – список вариантов трактовки словоформы (грамматически корректные разбиения на основу и флексию, неизменяемое слово).

Результат работы МОРФ1 (для знакомого слова) – список вариантов анализа, каждый из которых содержит: грамматические признаки словоформы и ссылку на словарную статью, описывающую семантическое значение слова.

Примеры:

стекла → (7 2 3 1 2) - существительное (неодуш., ср. род) *стекло* в форме: ед. число, родит. падеж
(7 2 3 2 (1 4)) - существительное (неодуш., ср. род) *стекло* в форме: мн. число, именит. или винит. падеж
(9 1 1 3 2 1 1) - глагол *стечь* в форме: прош. вр., женск. род, ед. число

Упрощенный вариант программы МОРФ1 – программа МОРФ3 – решает так называемую задачу ***лемматизации***: определяет только начальную форму слова, не формируя список грамматических характеристик словоформы.

Примеры:

стеки → стек, стечь

```
стекла   →   стекло, стечь
стеками  →   стек
```

Морфологический анализ незнакомых слов. Программа МОРФ2

На вход программы поступает сформированный МОРФ1 список вариантов трактовки словоформы.

Пример (словоформа *квазибиологом*)**:**

`квазибиологом+∅` (ср. *космодром/управдом*)
`квазибиолог+ом` (ср. *биолог+ом*)
`квазибиологом` (ср. *бегом*)

При обработке незнакомого слова МОРФ2 учитывает флексию и строение основы. В большинстве случаев исследование флексии не позволяет однозначно установить не только П-класс, род слов субстантивного склонения, вид спрягаемых слов, но даже М-класс анализируемого слова, так как, например, флексия *-а* встречается у слов всех четырех рассматриваемых М-классов (*стол-а*, *красив-а*, *дядин-а*, *ворош-а*). Для уточнения грамматических признаков незнакомых слов МОРФ2 учитывает следующие составляющие (диагностические сегменты) основы: префикс, суффикс или некоторую цепочку букв в конце основы, последнюю букву основы.

По префиксу можно обнаружить некоторые Н-слова и установить вид некоторых глаголов. Анализ суффикса помогает установить М-класс, П-класс, род (а иногда и одушевленность) слова субстантивного склонения, вид глагола или даже все нужные (описываемые в словарной статье) грамматические признаки слова. По последней букве основы легко уточняется П-класс, а иногда и М-класс слова. Программа МОРФ2 работает с таблицами, содержащими 28 префиксов и 67 суффиксов. Анализ незнакомого слова МОРФ2 начинает с варианта расщепления с максимальной длиной флексии.

Если анализируется не отдельно взятое слово, а слово в составе предложения, появляется возможность учета контекста (синтаксических связей данного слова с соседними). Информация о контексте передается программам морфологического анализа от объемлющих их программ синтаксического анализа с помощью предсказаний – списка ожидаемых грамматических признаков обрабатываемого слова. Так, при анализе незнакомого слова *Верхневартовск* в контексте *приехала из далекого Верхневартовска* ожидаемые характеристики последнего слова фрагмента таковы: неодушевленное существительное в форме единственного числа, родительного падежа.

В таких ситуациях результат работы МОРФ2 сопоставляется с предсказаниями, и, в случае соответствия, запоминается. Если же предсказание не подтвердилось, начинает обрабатываться другой вариант разбиения словоформы. Если ожидаемый результат не получен, либо слово признается неизменяемым, либо в нем ищутся и исправляются ошибки.

Для каждого отобранного варианта формируются результаты анализа словоформы (и вариант/варианты новой словарной статьи).

Пример (словоформа *квазибиологом*)**:**

`(7 0 1 1 (1 4))` - существительное (одуш. или неодуш., ср. род)

квазибиологом в форме: ед. число, именит. или винит. падеж
(7 1 1 1 5) - существительное (одуш., муж. род)
квазибиолог в форме: ед. число, творит. падеж
(11) – неизменяемое слово (возможно, наречие)

Заполнение словаря по грамматическим описаниям слов. Программа СЛОВ1

Основная сервисная программа автоматической генерации словарных статей – программа СЛОВ1. В ходе ее разработки были составлены таблицы соответствия словарной информации из словаря Зализняка и словарной информации ФМРС. Отметим, что программа СЛОВ1 автоматизирует трудоемкую, требующую хорошего знания ФМРС работу по составлению словарных статей. Действия, выполняемые программой, зачастую весьма нетривиальны из-за различий морфологической модели словаря Зализняка, и ФМРС. На вход программы поступает словарная статья, взятая из словаря Зализняка или (если такого слова там нет) сформированная экспертом.

Программа автоматически определяет: 1) основу записываемого в словарь системы слова; 2) номера М-класса, П-класса, С-класса; 3) наличие чередований и их контекст; 4) наличие других частных особенностей словоизменения. При работе с программой словарные статьи кодируются по определенным стандартным правилам, в частности, заменяются символы, отсутствующие на клавиатуре.

По первому элементу словарной информации из словаря Зализняка в большинстве случаев определяется номер М-класса, у слов субстантивного склонения также одушевленность и род, у спрягаемых слов – вид. Если, например, этот элемент «п», то слово относится к 8-му М-классу; «ж» – к 7-му М-классу, женскому роду, неодушевленное; «мо» – к 7-му М-классу, мужскому роду, одушевленное; «нсв» – к 9-му М-классу, несовершенному виду.

После определения М-класса происходит переход на соответствующую ветвь алгоритма, где по второму элементу определяется номер П-класса. Если второй элемент не цифра (это означает, что слово изменяется по необычной модели), СЛОВ1 фиксирует несовпадение номеров С-класса и М-класса (т.е. наличие соответствующего исключения) и формирует необходимый фрагмент словарной статьи.

Остальные элементы исходной словарной статьи либо уточняют номер П-класса, либо свидетельствуют о наличии в слове чередований, исключений или об отсутствии у слова некоторых форм. Например, символ «П2» означает, что у слова есть второй предложный падеж (локатив), символ «*» является признаком чередования. Для определения конкретного номера чередования СЛОВ1 анализирует строение начальной формы слова. Так, при обработке первого варианта слова *лев* номер чередования (4 – чередование: *ь - е*) определяется по буквам *ле*, стоящим перед последней согласной основы (буква *в* в данном случае неинформативна). Стандартный вариант основы (*льв-*) определяется по номерам П-класса и чередования.

Результатом работы программы СЛОВ1 является словарная статья или список таких словарных статей – в случае, когда слово из словаря Зализняка представляется в ФМРС семейством Н-слов и/или основ И-слов (для спрягаемых слов, например, программа строит словарную статью, описывающую личные формы глагола и деепричастия, и несколько статей для причастий).

Заполнение словаря по тексту. Программа СЛОВ2

Программа СЛОВ1 используется в ситуации, когда список слов, предназначенных для включения в компьютерный словарь, составлен заранее. Другая технологическая схема предполагает выявление незнакомых слов непосредственно в характерных текстах.

Режимы/сценарии работы с программой различаются:

- глубиной лингвистического анализа текста (пословный анализ, частичный синтаксический анализ, полный синтаксический анализ, синтактико-семантический анализ);

- «степенью самостоятельности» программ формирования словаря (работа без обращения за помощью к человеку, работа в диалоге с пользователем/администратором и под его контролем)

Морфологический синтез форм слова. Программа ФОРМ1

По словарной статье (знакомого слова) и набору значений ГП строится соответствующая словоформа.

Примеры:

```
ЛЕВ (животное), творит. падеж, ед. число  (7 0 0 1 5) → ЛЬВОМ
ЛЕВ (ден.единица), творит. падеж, ед. число     (7 0 0 1 5) → ЛЕВОМ
```

Морфологический синтез парадигмы. Программа ФОРМ2

По словарной статье (знакомого слова) строится массив всех форм этого слова. Порядок элементов массива определяется номером М-класса.

Примеры:

синтез всех форм знакомого существительного КАССИРША

```
КАССИРША        КАССИРШИ        - им.падеж,  ед. и мн.число
КАССИРШИ        КАССИРШ         - род.падеж, ед. и мн.число
КАССИРШЕ        КАССИРШАМ       - дат.падеж, ед. и мн.число
КАССИРШУ        КАССИРШ         - вин.падеж, ед. и мн.число
КАССИРШЕЙ       КАССИРШАМИ      - твор.падеж, ед. и мн.число
КАССИРШЕ        КАССИРШАХ       - предл.падеж,ед. и мн.число
```

синтез всех форм знакомого глагола ВОРОШИТЬ

```
ВОРОШИТЬ                        - начальная форма
ВОРОШИ    ВОРОШИТЕ              - формы повелит. наклонения
ВОРОШУ    (БУДУ ВОРОШИТЬ)       - 1 лицо, ед.ч, наст.и буд.вр.
ВОРОШИШЬ  (БУДЕШЬ ВОРОШИТЬ)     - 2 лицо, ед.ч, наст.и буд.вр.
ВОРОШИТ   (БУДЕТ ВОРОШИТЬ)      - 3 лицо, ед.ч, наст.и буд.вр.
ВОРОШИМ   (БУДЕМ ВОРОШИТЬ)      - 1 лицо, мн.ч, наст.и буд.вр.
ВОРОШИТЕ  (БУДЕТЕ ВОРОШИТЬ)     - 2 лицо, мн.ч, наст.и буд.вр.
ВОРОШАТ   (БУДУТ ВОРОШИТЬ)      - 3 лицо, мн.ч, наст.и буд.вр.
ВОРОШИЛ ВОРОШИЛА ВОРОШИЛО ВОРОШИЛИ  - формы прош.времени
ВОРОША    ВОРОШИВ               - деепричастия
```

Пусть написана управляющая программа, получающая на входе некоторую словоформу, обращающаяся к программе МОРФ1 (и – если слова нет в словаре – к МОРФ2) и генерирующая все формы (программа ФОРМ2) для каждого варианта анализа. Среди этих форм обязательно должна быть входная словоформа.

Примеры:

обработка незнакомого слова ХРЮША

```
ВАРИАНТ 1

склонение по образцу слова НОЖ/БОГАЧ
* значение ГП "одушевленность" неизвестно *
ХРЮШ                ХРЮШИ
ХРЮША           ХРЮШЕЙ
ХРЮШУ           ХРЮШАМ
ХРЮША / ХРЮШ    ХРЮШЕЙ / ХРЮШИ
ХРЮШОМ          ХРЮШАМИ
ХРЮШЕ           ХРЮШАХ

ВАРИАНТ 2

склонение по образцу слова МАРШ
* значение ГП "одушевленность" неизвестно *
ХРЮШ                ХРЮШИ
ХРЮША           ХРЮШЕЙ
ХРЮШУ           ХРЮШАМ
ХРЮША / ХРЮШ    ХРЮШЕЙ / ХРЮШИ
ХРЮШЕМ          ХРЮШАМИ
ХРЮШЕ           ХРЮШАХ

ВАРИАНТ 3

склонение по образцу слова ТУЧА/КАССИРША
* значение ГП "одушевленность" неизвестно *
ХРЮША           ХРЮШИ
ХРЮШИ           ХРЮШ
ХРЮШЕ           ХРЮШАМ
ХРЮШУ           ХРЮШ / ХРЮШИ
ХРЮШЕЙ          ХРЮШАМИ
ХРЮШЕ           ХРЮШАХ

ВАРИАНТ 4

склонение по образцу слова СВЕЖИЙ
ПОХРЮШЕЕ        ХРЮШЕЕ
ХРЮШ                ХРЮША           ХРЮШЕ           ХРЮШИ
ХРЮШИЙ              ХРЮШАЯ          ХРЮШЕЕ          ХРЮШИЕ
ХРЮШЕГО             ХРЮШЕЙ          ХРЮШЕГО         ХРЮШИХ
ХРЮШЕМУ             ЖРЮШЕЙ          ХРЮШЕМУ         ХРЮШИМ
ХРЮШЕГО&ХРЮШИЙ ХРЮШУЮ          ХРЮШЕЕ          ХРЮШИХ&ХРЮШИЕ
ХРЮШИМ              ХРЮШЕЙ          ХРЮШИМ          ХРЮШИМИ
ХРЮШЕМ              ХРЮШЕЙ          ХРЮШЕМ          ХРЮШИХ
```

```
ВАРИАНТ 5

спряжение по образцу слова ТОЧИТЬ/СЛЫШАТЬ
ХРЮШИТЬ
ХРЮШИ          ХРЮШИТЕ
ХРЮШУ          (БУДУ ХРЮШИТЬ)
ХРЮШИШЬ        (БУДЕШЬ ХРЮШИТЬ)
ХРЮШИТ         (БУДЕТ ХРЮШИТЬ)
ХРЮШИМ         (БУДЕМ ХРЮШИТЬ)
ХРЮШИТЕ        (БУДЕТЕ ХРЮШИТЬ)
ХРЮШАТ         (БУДУТ ХРЮШИТЬ)
ХРЮШИЛ    ХРЮШИЛА   ХРЮШИЛО   ХРЮШИЛИ
ХРЮША     ХРЮШИВ

ВАРИАНТ 6

неизменяемое слово типа АНТРАША
ХРЮША
```

Заметим, что если бы слово *хрюша* анализировалось с предсказаниями, результат был бы более точен. Так, при предсказании «существительное женского рода» был бы выдан только третий вариант, при предсказании «форма глагола – только пятый.

обработка незнакомого слова КРОВАТЬ

```
ВАРИАНТ 1

спряжение  по образцу слова ПИРОВАТЬ
* значение ГП "вид" неизвестно *
   (выбран несовершенный вид)
КРОВАТЬ
КРУЙ      КРУЙТЕ
КРУЮ      (БУДУ КРОВАТЬ)
КРУЕШЬ    (БУДЕШЬ КРОВАТЬ)
КРУЕТ     (БУДЕТ КРОВАТЬ)
КРУЕМ     (БУДЕМ КРОВАТЬ)
КРУЕТЕ    (БУДЕТЕ КРОВАТЬ)
КРУЮТ     (БУДУТ КРОВАТЬ)
КРОВАЛ    КРОВАЛА   КРОВАЛО   КРОВАЛИ
КРУЯ           КРОВАВ

ВАРИАНТ 2

склонение по образцу слова ПЕЧАТЬ
* значение ГП "одушевленность" неизвестно *
КРОВАТЬ        КРОВАТИ
КРОВАТИ        КРОВАТЕЙ
КРОВАТИ        КРОВАТЯМ
КРОВАТЬ        КРОВАТЕЙ / КРОВАТИ
КРОВАТЬЮ       КРОВАТЯМИ
КРОВАТИ        КРОВАТЯХ
```

ВАРИАНТ 3

неизменяемое слово типа ДЕСКАТЬ

Естественный язык и мышление. Гипотеза лингвистической относительности.

Вспомним некоторые базовые положения концепции интеллектуальной деятельности человека (мышления), которые подробно анализировались в курсе "Искусственный интеллект" (бакалавриат, 7 семестр).

Мышление (= Интеллект) – высшая форма Психического Отражения. Отражение по сфере сущностей, то есть **Понятийное Отражение.**
Понятие (об объекте) – Психическое Явление, отражающее ***сущность*** этого объекта.
Сущность – наиболее важные, глубинные характеристики предмета или явления, определяющие его свойства, поведение, развитие.
Понимание – выявление Сущности объекта.
Мышле'ние – способность человека к Понятийному Отражению.
Мы'шление – процесс Понятийного Отражения.

Известная триада психологии и педагогики: ***знания – умения – навыки***.
Знания – усвоенные Понятия.
Умения – способность выполнять новые действия в новых условиях.
Навыки – действия, автоматизировавшиеся в процессе их усвоения и выполнения.

Существует много различных теорий мышления, интеллекта.
Теория Интеллекта Жана Пиаже – одна из наиболее интересных теорий мышления и его развития.
Ж.Пиаже: но лишь один Интеллект "*тяготеет к тотальному равновесию, стремясь к тому, чтобы ассимилировать всю совокупность действительности и чтобы аккомодировать к ней действие, которое он освобождает от рабского подчинения изначальным 'здесь' и' теперь'*".

Для нас **ИНТЕЛЛЕКТ**:

- целенаправленное планирование поведения в меняющейся проблемной среде;
- перенос деятельности во внутренний план вместо выполнения поведенческих актов;
- работа с понятийными моделями среды и себя (на основе понятийного отражения);
- скоординированная совокупность мыслительных/интеллектуальных операций – как абстрактных (метод рассуждения по аналогии), так и конкретных (способ решения определенного типа задач);

В контексте нашего курса мы обратим внимание на связь мышления с естественными языками человека и на возможность влияния того языка (которым владеет/который является родным для пользователя) на его взаимодействия с компьютером.
Об одном ошибочном в целом (но получившем большой социальный резонанс) взгляде на связь языка и мышления замечательно написал в свое время наш выдающийся лингвист Владимир Андреевич Звегинцев.

В.А. Звягинцев Теоретико-лингвистические предпосылки гипотезы Сепира-Уорфа
НОВОЕ В ЛИНГВИСТИКЕ – М.: Иностранная литература, 1960.

[HTML]http://www.classes.ru/grammar/148.new-in-linguistics-1/source/worddocuments/_15.htm (05.02.2020).

"Проблема взаимоотношений языка и мышления является традиционной для науки о языке и уходит своими корнями в классическую древность. Первоначально и преимущественно эта проблема рассматривалась в направлении влияния категорий мышления на становление языковых и в первую очередь грамматических категорий.

. . .

В пределах самого языкознания это второе направление в исследовании проблемы языка и мышления было начато гениальным основоположником общего языкознания и философии языка (в ее идеалистическом толковании) В. Гумбольдтом. Его постановка вопроса представлена в следующем высказывании: «Так как ко всякому объективному восприятию неизбежно примешивается субъективное, то каждую человеческую индивидуальность независимо от языка можно считать носителем особого мировоззрения. Само его образование осуществляется через посредство языка, поскольку слово в противоположность душе превращается в объект всегда с примесью собственного значения и таким образом привносит новое своеобразие. Но в этом своеобразии, так же как и в речевых звуках, в пределах одного языка наблюдается всепроникающая тождественность, а так как к тому же на язык одного народа воздействует однородное субъективное начало, то в каждом языке оказывается заложенным свое мировоззрение. Если звук стоит между предметом и человеком, то весь язык в целом находится между человеком и воздействующей на него внутренним и внешним образом природой. Человек окружает себя миром звуков, чтобы воспринять и усвоить мир предметов. Это положение ни в коем случае не выходит за пределы очевидной истины. Так как восприятие и деятельность человека зависят от его представлений, то его отношение к предметам целиком обусловлено языком. Тем же самым актом, посредством которого он создает язык, человек отдает себя в его власть: каждый язык описывает вокруг народа, которому он принадлежит, круг, из пределов которого можно выйти только в том случае, если вступаешь в другой круг» (Л. Витгенштейн "Логико-философский трактат", М., 1958, стр. 45.).

На совершенно аналогичные мысли (причем, несомненно, независимо от европейской научной традиции) натолкнуло американских исследователей изучение культуры и языка первых обитателей американского континента — многочисленных индейских племен. Формы культуры, обычаи, этические и религиозные представления, с одной стороны, и структура языков—с другой, имели у американских индейцев чрезвычайно своеобразный характер и резко отличались от всего того, с чем до знакомства с ними приходилось сталкиваться в этих областях ученым. Это обстоятельство и подсказало американским ученым мысль о прямой связи между формами языка, культуры и мышления.

Более четкую и определенную форму этим мыслям придал один из самых талантливых представителей американской науки о языке — Эдуард Сепир. Его высказывание по данному вопросу Б. Уорф избрал в качестве эпиграфа к своей основной статье. Оно многократно повторяется и во многих других работах, посвященных интерпретации идей Б. Уорфа, но его никак нельзя опустить и здесь, так как оно открывает доступ к пониманию всей концепции Б. Уорфа. «Люди живут не только в объективном мире

вещей,— пишет Э. Сепир,— и не только в мире общественной деятельности, как это обычно полагают; они в значительной мере находятся под влиянием того конкретного языка, который является средством общения для данного общества. Было бы ошибочным полагать, что мы можем полностью осознать действительность, не прибегая к помощи языка, или что язык является побочным средством разрешения некоторых частных проблем общения и мышления. На самом же деле «реальный мир» в значительной степени бессознательно строится на основе языковых норм данной группы... Мы видим, слышим и воспринимаем так или иначе те или другие явления главным образом благодаря тому, что языковые нормы нашего общества предполагают данную форму выражения».

То, что для Э. Сепира было одной из многих проблем, которыми он занимался, составило содержание всего научного творчества Б. Уорфа. Специфичность его творчества не только в том, что он попытался общую формулу Э. Сепира наполнить конкретным содержанием и приложить ее к изучению собственно языкового материала, но также и в том (и это в большей мере), что он развил ее, расширил, придав ей форму своеобразной метафизической системы, а главное, сделал из нее крайние логические выводы. Это, кстати говоря, тотчас же обнаружило все слабые стороны данной идеи, получившей название гипотезы Сепира—Уорфа или теории лингвистической относительности". [Выделено бирюзовым фоном лектором]

На формирование взглядов Б. Уорфа о влиянии языка на деятельность человека большое влияние оказала его профессия (в отличие от Э. Сэпира он был по образованию не лингвистом, а химиком-технологом, окончившим MIT и работавшим в страховой компании).

Б. Л. Уорф Отношение норм поведения и мышления к языку
НОВОЕ В ЛИНГВИСТИКЕ – М.: Иностранная литература, 1960.
[HTML]http://www.classes.ru/grammar/148.new-in-linguistics-1/source/worddocuments/_16.htm (05.02.2020).

> «Люди живут не только в объективном мире вещей и не только в мире общественной деятельности, как это обычно полагают; они в значительной мере находятся под влиянием того конкретного языка, который является средством общения для данного общества. Было бы ошибочным полагать, что мы можем полностью осознать действительность, не прибегая к помощи языка, или что язык является побочным средством разрешений некоторых частных проблем общения и мышления. На самом же деле «реальный мир» в значительной степени бессознательно строится на основе языковых норм данной группы... Мы видим, слышим и воспринимаем так или иначе те или другие явления главным образом благодаря тому, что языковые нормы нашего общества предполагают данную форму выражения».
>
> Эдуард Сепир

"Вероятно, большинство людей согласится с утверждением, что принятые нормы употребления слов определяют некоторые формы мышления и поведения; однако это

предположение обычно не идет дальше признания гипнотической силы философского и научного языка, с одной стороны, и модных словечек и лозунгов — с другой.

Ограничиться только этим — значит не понимать сути одной из важнейших форм связи, которую Сепир усматривал между языком, культурой и психологией и которая кратко сформулирована в приведенной выше цитате.

Мы должны признать влияние языка на различные виды деятельности людей не столько в особых случаях употребления языка, сколько в его постоянно действующих общих законах и в его повседневной оценке им тех или иных явлений.

ОБОЗНАЧЕНИЕ ЯВЛЕНИЯ И ЕГО ВЛИЯНИЕ НА ДЕЙСТВИЯ ЛЮДЕЙ

Я столкнулся с одной из сторон этой проблемы еще до того, как начал изучать Сепира, в области, обычно считающейся очень отдаленной от лингвистики. Это произошло во время моей работы в обществе страхования от огня. В мои задачи входил анализ сотен докладов об обстоятельствах, приведших к возникновению пожара или взрыва. Я фиксировал чисто физические причины, такие, как неисправная проводка, наличие или отсутствие воздушного пространства между дымоходами и деревянными частями зданий и т. п., и результаты обследования описывал в соответствующих терминах. При этом я не ставил перед собой никакой другой задачи. Но с течением времени стало ясно, что не только сами физические обстоятельства, но и обозначение этих обстоятельств было иногда тем фактором, который, через поведение людей, являлся причиной пожара. Этот фактор обозначения становился яснее всего тогда, когда это было языковое обозначение, исходящее из названия, или обычное описание подобных обстоятельств средствами языка.

Так, например, около склада так называемых *gasoline drums* (бензиновых цистерн) люди ведут себя определенным образом, т. е. с большой осторожностью; в то же время рядом со складом с названием *empty gasoline drums* (пустые бензиновые цистерны) люди ведут себя иначе — недостаточно осторожно, курят и даже бросают окурки. Однако эти «пустые» (*empty*) цистерны могут быть более опасными, так как в них содержатся взрывчатые испарения. При наличии реально опасной ситуации лингвистический анализ ориентируется на слово «пустой», предполагающее отсутствие всякого риска. Существуют два различных случая употребления слова *empty*: 1) как точный синоним слов — *null, void, negative, inert* (порожний, бессодержательный, бессмысленный, ничтожный, вялый) и 2) в применении к обозначению физической ситуации, не принимая во внимание наличия паров, капель жидкости или любых других остатков в цистерне или другом вместилище. Обстоятельства описываются с помощью второго случая, а люди ведут себя в этих обстоятельствах, имея в виду первый случай. Это становится общей формулой неосторожного поведения людей, обусловленного чисто лингвистическими факторами.

На лесохимическом заводе металлические дистилляторы были изолированы смесью, приготовленной из известняка, именовавшегося на заводе «центрифугированным известняком». Никаких мер по предохранению этой изоляции от перегревания и соприкосновения с огнем принято не было. После того как дистилляторы были в употреблении некоторое время, пламя под одним из них проникло к известняку,

который, ко всеобщему удивлению, начал сильно гореть. Поступление испарений уксусной кислоты из дистилляторов способствовало превращению части известняка в ацетат кальция. Последний при нагревании огнем разлагается, образуя воспламеняющийся ацетон. Люди, допускавшие соприкосновение огня с изоляцией, действовали так потому, что само название «известняк» (*limestone*) связывалось в их сознании с понятием *stone* (камень), который «не горит».

Огромный железный котел для варки олифы оказался перегретым до температуры, при которой он мог воспламениться. Рабочий сдвинул его с огня и откатил на некоторое расстояние, но не прикрыл. Приблизительно через одну минуту олифа воспламенилась. В этом случае языковое влияние оказалось более сложным благодаря переносу значения (о чем ниже будет сказано более подробно) «причины» в виде контакта или пространственного соприкосновения предметов на понимание положения *on the fire* (на огне) в противоположность *off the fire* (вне огня). На самом же деле та стадия, когда наружное пламя являлось главным фактором, закончилась; перегревание стало внутренним процессом конвенции в олифе благодаря сильно нагретому котлу и продолжалось, когда котел был уже вне огня (*off the fire*).

Электрический рефлектор, висевший на стене, мало употреблялся и одному из рабочих служил удобной вешалкой для пальто. Ночью дежурный вошел и повернул выключатель, мысленно обозначая свое действие как *turning on the light* (включение света). Свет не загорелся, и это он мысленно обозначил как *light is burned out* (перегорели пробки). Он не мог увидеть свечения рефлектора только из-за того, что на нем висело старое пальто. Вскоре пальто загорелось от рефлектора, отчего вспыхнул пожар во всем здании.

Кожевенный завод спускал сточную воду, содержавшую органические остатки, в наружный отстойный резервуар, наполовину закрытый деревянным настилом, а наполовину открытый. Такая ситуация может быть обозначена как *pool of water* (резервуар, наполненный водой). Случилось, что рабочий зажигал рядом паяльную лампу и бросил спичку в воду. Но при разложении органических остатков выделялся газ, скапливавшийся под деревянным настилом, так что вся установка была отнюдь не *watery* (водной). Моментальная вспышка огня воспламенила дерево и очень быстро распространилась на соседнее здание.

Сушильня для кож была устроена с воздуходувкой в одном конце комнаты, чтобы направить поток воздуха вдоль комнаты и далее наружу через отверстие на другом конце. Огонь возник в воздуходувке, которая направила его прямо на кожи и распространила искры по всей комнате, уничтожив таким образом весь материал. Опасная ситуация создалась таким образом благодаря термину blower (воздуходувка), который является языковым эквивалентом *that which blows* (то, что дует), указывающим на то, что основная функция этого прибора — *blow* (дуть). Эта же функция может быть обозначена как *blowing air for drying* (раздувать воздух для просушки); при этом не принимается во внимание, что он может «раздувать» и другое, например искры и языки пламени. В действительности воздуходувка просто создает поток воздуха и может втягивать воздух так же, как и выдувать. Она должна была быть поставлена на другом конце помещения, там, где было отверстие, где она могла бы тянуть воздух над шкурами, а затем выдувать его наружу.

Рядом с тигелем для плавки свинца, имевшим угольную топку, была помещена груда *scrap lead* (свинцового лома) — обозначение, вводящее в заблуждение, так как на самом деле «лом» состоял из листов старых радиоконденсаторов, между которыми все еще были парафиновые прокладки. Вскоре парафин загорелся и поджег крышу, половина которой была уничтожена.

Количество подобных примеров может быть бесконечно увеличено. Они показывают достаточно убедительно, как рассмотрение лингвистических формул, обозначающих данную ситуацию, может явиться ключом к объяснению тех или иных поступков людей и каким образом эти формулы могут анализироваться, классифицироваться и соотноситься в том мире, который «в значительной степени бессознательно строится на основании языковых норм данной группы». Мы ведь всегда исходим из того, что язык лучше, чем это на самом деле имеет место, отражает действительность".

В той же работе Б.Уорф отмечает:

"ньютоновские понятия пространства, времени и материи не есть данные интуиции. Они даны культурой и языком. Именно из этих источников и взял их Ньютон".

В более поздних работах подобные Б. Уорф переходит к более высокому уровню абстракции и говорит уже собственно о "лингвистической относительности".

Бенджамен Л. Уорф Наука и языкознание
НОВОЕ В ЛИНГВИСТИКЕ – М.: Иностранная литература, 1960.
[HTML]http://www.classes.ru/grammar/148.new-in-linguistics-1/source/worddocuments/15.htm (05.02.2020).

"Когда лингвисты смогли научно и критически исследовать большое число языков, совершенно различных по своему строю, их опыт обогатился, основа для сравнения расширилась, они столкнулись с нарушением тех закономерностей, которые до того считались универсальными, и познакомились с совершенно новыми типами явлений. Было установлено, что основа языковой системы любого языка (иными словами, грамматика) не есть просто инструмент для воспроизведения мыслей. Напротив, грамматика сама формирует мысль, является программой и руководством мыслительной деятельности индивидуума, средством анализа его впечатлений и их синтеза. Формирование мыслей — это не независимый процесс, строго рациональный в старом смысле этого слова, но часть грамматики того или иного языка и различается у различных народов в одних случаях незначительно, в других — весьма существенно, так же как грамматический строй соответствующих языков. Мы расчленяем природу в направлении, подсказанном нашим родным языком. Мы выделяем в мире явлений те или иные категории и типы совсем не потому, что они (эти категории и типы) самоочевидны; напротив, мир предстает перед нами как калейдоскопический поток впечатлений, который должен быть организован нашим сознанием, а это значит в основном — языковой системой, хранящейся в нашем сознании. Мы расчленяем мир. организуем его в понятия и распределяем значения так, а не иначе в основном потому, что мы — участники соглашения, предписывающего подобную систематизацию. Это соглашение имеет силу для определенного речевого коллектива и закреплено в системе моделей нашего языка. Это соглашение, разумеется, никак и никем не сформулировано и лишь подразумевается, и тем не менее мы — участники этого

соглашения] мы вообще не сможем говорить, если только не подпишемся под систематизацией и классификацией материала, обусловленной указанным соглашением.

Это обстоятельство имеет исключительно важное значение для современной науки, поскольку из него следует, что никто не волен описывать природу абсолютно независимо, но все мы связаны с определенными способами интерпретации даже тогда, когда считаем себя наиболее свободными. Человеком, более свободным в этом отношении, чем другие, оказался бы лингвист, знакомый со множеством самых разнообразных языковых систем. Однако до сих пор таких лингвистов не было. Мы сталкиваемся, таким образом, с **новым принципом относительности, который гласит, что сходные физические явления позволяют создать сходную картину вселенной только при сходстве или по крайней мере при соотносительности языковых систем".**
[Красным шрифтом текст выделил лектор – М.Г.Мальковский]

Как отмечал В.А. Звягинцев, существенное влияние на развитие американской лингвистики оказало изучение языков американских индейцев:
"Формы культуры, обычаи, этические и религиозные представления, с одной стороны, и структура языков—с другой, имели у американских индейцев чрезвычайно своеобразный характер и резко отличались от всего того, с чем до знакомства с ними приходилось сталкиваться в этих областях ученым. Это обстоятельство и подсказало американским ученым мысль о прямой связи между формами языка, культуры и мышления". (см. выше).

В другой своей работе Б.Уорф приводит весьма любопытные примеры различий в английском языке и языке индейцев Шауни.

Бенджамен Л. Уорф Лингвистика и логика
НОВОЕ В ЛИНГВИСТИКЕ – М.: Иностранная литература, 1960.
[HTML] http://www.classes.ru/grammar/148.new-in-linguistics-1/source/worddocuments/16.htm (05.02.2020.)

"В английском языке предложения ***I pull the branch aside*** «Я отодвигаю ветку» и ***I have an extra toe on my foot*** «У меня лишний палец на ноге» мало чем похожи друг на друга. Можно даже сказать, что в них нет ничего общего, за исключением местоимения в функции подлежащего и настоящего времени глаголов, являющихся общими в этих предложениях, согласно правилам английского синтаксиса. С обывательской и даже с научной точки зрения эти предложения различны, так как они повествуют о вещах, существенно отличающихся друг от друга. Таков довод «Всякого человека», обладающего естественным логическим мышлением. Формальная логика старого типа, вероятно, поддержала бы его.

Если, далее, мы обратимся к беспристрастному научно мыслящему наблюдателю, говорящему по-английски, и попросим его произвести анализ данных предложений и посмотреть, не пропустили ли мы каких-либо черт сходства, он почти наверное подтвердит то, что сказали «Всякий человек» и логик. Человек, которого мы попросили проанализировать наш случай, возможно, не будет смотреть глазами логика старой школы и с удовольствием уличит последнего в ошибке. Но все-таки ему придется с

грустью признать, что это ему не удалось. «Я бы очень хотел сделать вам приятное,— скажет он,— но, сколько я ни пытался, я не могу обнаружить никакого сходства между этими двумя явлениями».

К этому времени в нас возникает своего рода упрямство: нам становится интересно, нашел ли бы марсианин еще какое-нибудь сходство между нашими предложениями? И оказывается, что с точки зрения лингвиста вовсе не надо отправляться так далеко. Мы еще не обыскали нашу планету, чтобы выяснить, во всех ли языках эти два утверждения так же несравнимы, как в нашей речи Оказывается, что в языке шауни оба утверждения последовательно выглядят так: ***ni-Г Hawa-'ko-n-a*** и ***ni-Г Oawa-'ko-ftite*** (***H*** здесь обозначает ***th***, как в ***thin***, а апостроф обозначает перерыв дыхания). Оба предложения имеют большое сходство, практически они различаются только в последней своей части. Более того, в шауни начало предложения обычно является основной, самой важной частью. Оба предложения начинаются с *ni-*(«*I*»), которое фактически является приставкой. Далее идет действительно важная часть — ключевое слово *l'Hawa* — обычный для шауни термин, обозначающий вилообразный предмет...

О следующем элементе ***-'ко*** мы не можем сказать ничего определенного, кроме того, что он согласуется по форме с одним из вариантов суффикса ***-a'kw*** или ***-a'ко***, обозначающим дерево, куст, часть дерева, ветку и т. п. В первом предложении ***-п-*** обозначает ***by hand action*** «посредством действия руки» и может быть или непосредственной причиной основного состояния (вилообразной формы), или его дальнейшим преобразованием, или же соединять оба эти понятия. Конечное ***-a*** означает, что субъект (****Г***) производит это действие по отношению к соответствующему предмету.

Таким образом, первое предложение значит: ***I pull it*** (что-то подобное ветке дерева) ***more open or apart where it forks*** «Я отодвинул это дальше от места развилки». В другом предложении суффикс ***-0ite*** означает ***pertaining to the toes*** «принадлежащий пальцам», а отсутствие других суффиксов указывает на то, что субъект говорит о состоянии своего собственного тела. Поэтому предложение может означать только: ***I have an extra toe forking out like a branch from anormal toe*** «У меня лишний палец, ответвляющийся от нормального пальца, как ветка дерева»".

Действительно в различных естественных языках существуют разные способы выражения одного и того же содержания.

Простейший пример: в русском языке глагол несовершенного вида в форме настоящего времени (например, *работает*) может обозначать либо действие, выполняемое в настоящий момент времени (*генератор сейчас работает*), либо действие выполняемое иногда (*генератор работает при низком напряжении в сети*) или всегда (*генератор работает на нашей электростанции уже 10 лет без перерыва*).

В английском языке в первом случае мы употребляем глагол во времени **Simple** (или Indefinite – простой или неопределенный: *works*), во втором – Progressive (или Continuous – прогрессивный, продолженный или длительный: *is working*). В третьем случае придется воспользоваться обстоятельственными конструкциями.

Оценивая гипотезу лингвистической относительности в целом, отметим, что ее авторы отметив (и исследовав) влияние языка на человеческую деятельность, потеряли из

виду/забыли об обратном и первичном процессе формирования и самого возникновения языка в процессе человеческой деятельности (в филогенезе), а сам факт влияния языка на деятельность человека довели до абсурда. Носители двух сильно отличающихся друг от друга естественных языков (изучивших их в процессе своего индивидуального развития – онтогенезе) воспринимают те или иные явления иначе не просто потому, что говорят на разных языках, но и (в первую очередь потому), что эти разные языки возникли в разных социальных и культурных контекстах.

Очевидные и широко известные факты о наличии в языках северных народов десятков слов, обозначающих различные видов/состояний снега подтверждают факт зависимости естественных языков от условий их возникновения. Высказывание Б.Уорфа:

"То, что американские индейцы, владеющие только своими родными языками, никогда не выступали в качестве ученых или исследователей, не имеет отношения к делу. Игнорировать свидетельство своеобразия человеческого разума, которое предоставляют их языки, — это все равно, что ожидать от ботаников исчерпывающего описания растительного мира, зная, что они изучили только растения, употребляемые для пищи, и оранжерейные розы". — доказывает не мифическую лингвистическую относительность, а именно факт влияния деятельного контекста на формирование языка.

В наше время "сильная" версия гипотезы лингвистической относительности ("язык определяет то, как мы воспринимаем мир") отвергается и научным сообществом, и повседневным речевой деятельностью рядового носителя языка. Вряд ли кто-то верит, что "существующие в сознании человека системы понятий, а, следовательно, и существенные особенности его мышления определяются тем конкретным языком, носителем которого этот человек является".

Энциклопедия Кругосвет. Универсальная научно-популярная онлайн-энциклопедия. "Лингвистической относительности гипотеза"
[HTML]
http://www.krugosvet.ru/enc/gumanitarnye_nauki/lingvistika/LINGVISTICHESKO_OTNOSITELNOSTI_GIPOTEZA.html (05.02.2020.)

В то же время авторы концепции ввели в круг лингвистических исследований связь языка с культурой (этнолингвистика). В "Словаре культуры XX века" В.П.Руднёв отмечает, что гипотеза лингвистической относительности: "сыграла большую роль в культуре XX в., но не столько в академической лингвистике, которая к ней относилась с подозрением, сколько в смежных областях, в аналитической философии, в междисциплинарных культурологических исследованиях". Он же отмечает давно известный факт: ". . . случай в лондонском метро, о котором когда-то писали газеты. Таблички на дверях, гласившие "Выхода нет", по совету социологов заменили табличками "Выход рядом", что на несколько процентов понизило число самоубийств в Лондоне". (показывающий, необходимость аккуратного отношения к неоднозначным конструкциям языка; явления такого рода относятся, пожалуй, не к социолингвистике, а к психолингвистике).

Словари и энциклопедии на Академике

https://dic.academic.ru/dic.nsf/enc_culture/831/%D0%93%D0%98%D0%9F%D0%9E%D0%A2%D0%95%D0%97%D0%90 (05.02.2020.)

Серьезные научные исследования в области социо- и этнолингвистики показывают, что если естественный язык и оказывает какое-либо влияние на поведение человека, то проявляется это влияние не при формировании модели мира, а в гораздо менее масштабных и специфических ситуациях.
В этой связи интерес представляет работа:

Сьюзэн М. Эрвин Семантический сдвиг при двуязычии
НОВОЕ В ЛИНГВИСТИКЕ, вып. 6 – М.: Прогресс, 1972.
[HTML]http://www.classes.ru/grammar/153.new-in-linguistics-6/source/worddocuments/_23.htm (05.02.2020.)
Автор демонстрирует четкость и научность при выборе методики эксперимента, описании и анализе полученных данных.

В основе исследования лежат наблюдения Робертса и Леннеберга (E.H. Lenneberg and J.H. Roserts, The Language of Experience, international Journal of American Linguistics, Memoir 13, 1956, стр. 22.), заметивших, что у индейцев зуни, знающих два языка (зуни и английский) система обозначения цветов иная, чем у одноязычных зуни.

Объект исследования – семантические сдвиги при двуязычии (владении двумя языками). Автором предложен метод, "позволяющий предсказывать систему названий цветов у двуязычных носителей; этот метод основан на несложной теории «словесной медиации» (verbal mediation)".

А одна из целей – "объяснить явление семантической интерференции, или сдвига в значении слов, возникающего под влиянием второго языка".
"Участниками экспериментов были индейцы-навахи, живущие в резервации, в том числе 28 одноязычных, 21 двуязычный с доминирующим языком английским и 13 двуязычных с доминирующим языком навахо; в большинстве это были женщины в возрасте от 17 до 70 лет".
Испытуемым предъявлялись цветные фишки из стандартного тестового набора (100 цветов).
"Цвета предъявлялись вперемешку, но так, чтобы сложные оттенки не шли подряд. Сначала испытуемый должен был называть цвета на языке навахо; а несколько позже, но в этот же день — по-английски. Испытуемых просили называть цвета непринужденно, как будто в разговоре с другом".

"В ходе эксперимента записывались произносившиеся названия цветов и время реакции. Если испытуемый не успевал подобрать название цвета за 30 сек., фишку убирали, а потом она предъявлялась повторно в ряду с оставшимися фишками".
В ходе экспериментов были отмечены интересные особенности при назывании цветов фишек:

"Желтый цвет. Подбирать названия оттенкам в желтой части спектра по-английски оказалось труднее, чем на языке навахо. Слово языка навахо Litso покрывает более широкий отрезок спектра и в своем центральном значении имеет гораздо более высокую вероятность употребления, чем англ. yellow. Центральным для обоих языков был оттенок № 16, но из навахов его назвали словом Litso 89%, а из английской группы словом yellow его обозначили только 34%. Оттенок № 16 слабее насыщен, чем

«хороший» желтый цвет, и были испытуемые, которые назвали его словами tan «желто-коричневый», beige «бежевый», green «зеленый» и brown «коричневый». Запаздывание при назывании этого оттенка у английских одноязычных было наибольшим, если не считать оттенок № 84. Их реакция была значительно медленней, чем у группы навахо ($p < 0{,}0001$). Таким образом, для оттенка № 16 можно было предсказать, что двуязычные, говоря по-английски, сначала внутренне назовут цвет словом навахо, и это повысит вероятность употребления ими англ. yellow".

При анализе полученных экспериментальных данных Сьюзэн М. Эрвин отмечает: "Данные эксперимента подтверждают, что в ряде ситуаций правомерно описывать процессы называния у двуязычных в терминах скрытых реакций, влияющих на речь через перевод. Сформулируем эти ситуации в более общем виде.

(1) Когда в одном языке имелось высокоупотребительное название, а в другом языке такого названия не было, в речи двуязычных на том и на другом языке преобладало это высоковероятное название и, соответственно, его переводной эквивалент.

(2) Если языки проводили границу между двумя категориями, имеющими названия в каждом языке, в разных точках, то двуязычные, определяя эту граничную точку, ориентировались на свой доминирующий язык.

(3) Если категория одного языка обнимала область значений, которой во втором языке соответствует две категории, то граничная точка между этими последними варьировала и зависела от степени владения вторым языком.

(4) В случае, когда область, покрываемая в одном языке двумя категориями, другим языком членилась на три категории, так что граничная точка между категориями первого языка оказывалась внутри средней категорий второго, двуязычные при речи на втором языке сужали эту среднюю категорию.

Первоначально мы предполагали, что в распоряжении обеих групп двуязычных имеются примерно одинаковые по богатству наборы названий цветов и что ожидаемые различия в реакциях будут связаны с разницей в силе давления, оказываемого на каждую из групп ее доминирующим языком — навахо или английским. Это предположение оказалось ошибочным. Результаты эксперимента свидетельствуют, что с усвоением английского языка словарь названий цветов становится богаче, причем наблюдается зависимость между показателем степени доминантности языка и вероятностью употребления таких слов, как lavender «бледно-лиловый» и violet «лиловый» ($p < < 0{,}005$ по t-критерию), имеющих в английском языке меньшую частотность, чем purple «фиолетовый» п.

Подобные различия наблюдались и при речи двуязычных на языке навахо: запас названий цветов у пожилых испытуемых был богаче. Употребительность как нав. tatLqid, так и нав. Liba зависела от возраста испытуемых, хотя, в общем, Liba употреблено большим числом двуязычных. Некоторые молодые двуязычные с доминантностью английского пользовались этими словами неправильно с точки зрения одноязычной нормы. Например, один из них называл словом tatLqid только оттенки № 18 и № 63 и ни один из промежуточных оттенков. Таким образом, оказывается, что вероятность переводных соответствий между

словами двух языков не всегда можно точно предсказать из-за неодинаковой степени владения словарем.

Вероятно, процессы, аналогичные выявленным при назывании цветов, будут иметь место и при семантических сдвигах в других областях значений, например, у слов, относящихся к эмоциональной сфере. Остается неясным, приложимо ли то же простое объяснение и к сдвигам в значении слов, относящихся к дискретным категориям, а не к сплошному спектру свойств.

Полученные результаты не дают оснований предполагать какие-либо различия в цветовом видении у одноязычных и двуязычных испытуемых непосредственно в момент восприятия. Опыты Леннеберга с языком зуни не обнаружили различий в восприятии цвета у носителей языка зуни и английского при одновременной демонстрации цветовых образцов.

Как показали тесты на сортировку цветов и опыты по определению пороговых значений, языковые категории не оказывают влияния на остроту цветового восприятия. В то же время очевидно, что в ситуациях, где можно предположить «словесную медиацию» (verbal mediation), одноязычные и двуязычные испытуемые будут вести себя неодинаково".

Эксперименты Сьюзэн М. Эрвин ни в коей мере не служат подтверждением гипотезы лингвистической относительности. Однако они показывают, что при разработке пользовательского интерфейса, **предполагающего "словесную медитацию"** (в зависимости от цвета фона должна последовать та или иная реакция пользователя), следует принимать во внимание и социолингвистические факторы.

Раздел 3. Естественный язык в человеко-машинном общении.

ОБЩЕНИЕ ЧЕЛОВЕКА С СИСТЕМОЙ ИИ

(Искусственный интеллект и естественный язык)

Естественный язык в работах по искусственному интеллекту. Первые шаги.

В 60-е – 70-е гг. XX века специалистов в области искусственного интеллекта интересовала прежде всего проблема моделирования понимания сообщений/текстов. Были реализованы системы, способные «поддерживать диалог» с человеком на естественном языке, интерпретировать словесные команды роботу (действующему в «Мире кубиков»), решать задачи в словесной формулировке.

Система ELIZA (DOCTOR) – Дж. Вейзенбаум (1966)

Система относится к классу Conversation Machines, разработчики которых надеялись на то, что их творения смогут пройти ***Тест Тьюринга.***
Версия **DOCTOR** моделирует диалог врача-психоаналитика (система) с пациентом (человек).

Фрагмент диалога с системой ELIZA:

USER: MY MOTHER TAKES CARE OF ME

ELIZA: WHO ELSE IN YOUR FAMILY TAKES CARE OF YOU

USER: MY FATHER

ELIZA: YOUR FATHER

USER: YOU ARE LIKE MY FATHER IN SOME WAYS

ELIZA: WHAT RESEMBLANCE DO YOU SEE

USER: YOU ARE NOT VERY AGGRESSIVE

ELIZA: WHAT MAKES YOU THINK I AM NOT VERY AGGRESSIVE

Для формирования ответов система использовала простейшие шаблоны и связанные с ними правила, например:

(<any-1> you <any-3> me) → (what makes you think I <any-3> you)

USER: IT SEEMS THAT YOU HATE ME

ELIZA: WHAT MAKES YOU THINK I HATE YOU

(<any-1> you <any-3>) -> (what makes you think I <any-3>)

USER: YOU ARE NOT VERY AGGRESSIVE

ELIZA: WHAT MAKES YOU THINK I AM NOT VERY AGGRESSIVE

Система APRIL – М.Г.Мальковский (1969)

Система APRIL решает арифметические задачи в словесной формулировке. Результатом ее работы является ответ (формула или число), иногда сопровождаемый некоторыми комментариями, или сообщение (также сопровождаемое комментариями) о том, что задачу решить не удалось.

Система решает довольно узкий круг задач из школьного курса математики (1-2-й классы) – так называемые задачи «на владение». Каждая задача решается независимо от других.

На входной язык системы наложены достаточно жесткие ограничения, которые, впрочем, абсолютно естественны для рассматриваемого класса задач. Так, условие задачи должно быть сформулировано в виде трех простых предложений, а все слова, встречающиеся в тексте условия, должны быть знакомы системе (записаны в ее словарь).

Наиболее серьезное внимание было уделено разработке методов, позволяющих учитывать возможность варьирования словесных формулировок условий задач и употребления в тексте неполных (эллиптичных) конструкций. Так, внешне непохожие задачи № 1, № 2 и № 3 система сводит к одному типу задач, а решая задачи, которые человек воспринимает как имеющие решение и притом единственное, несмотря на отсутствие подразумеваемых уточняющих слов, она делает примерно те же предположения, что и человек.

Задача N 1: У Саши 5 яблок. У Пети 10 груш. Сколько фруктов у ребят?

ОТВЕТ: 15

Задача N 2: У человека 10 книг. У второго человека в 2 раза меньше. Сколько книг у них вместе?

ОТВЕТ: 15

Задача N 3: На стол положили 6 книг. Затем еще 4 книги. Сколько книг на столе?

ЕСЛИ ВО ВСЕХ ФРАЗАХ ИМЕЕТСЯ В ВИДУ "СТОЛ", ТО
ЕСЛИ В НАЧАЛЬНЫЙ МОМЕНТ КОЛИЧЕСТВО ОБЪЕКТОВ "КНИГА" В МЕСТЕ "СТОЛ" = 0,ТО
ПРЕДПОЛОЖИТЕЛЬНЫЙ ОТВЕТ: 10

Первые эксперименты показали и принципиальную возможность моделирования процессов владения языком, и чрезвычайную сложность естественного языка как объекта исследования.

Общение человека с компьютером (естественный язык и естественность общения)

Наиболее существенными и привлекательными (в контексте задачи общения с компьютером / компьютерной системой) свойствами ЕЯ являются:

- максимально широкое **использование его человеком в своей повседневной деятельности** (это избавляет от необходимости специального изучения формализованного языка общения с ЭВМ и от трудностей, связанных с формулировкой заданий и запросов на таком языке);

- возможность использования естественного языка для **выражения качественно различного содержания** с любой доступной или желательной человеку степенью строгости и полноты (что гарантирует чрезвычайную широту сферы общения – как в плане охвата самых разнообразных предметных областей, так и в плане варьирования формулировок);

- его **открытость** и способность служить метаязыком для самого себя (что обеспечивает расширяемость используемых языковых средств).

Эти обстоятельства (обычно упоминается первое – не только потому, что оно действительно важно, но и потому, что оно абсолютно очевидно, лежит на поверхности) служат очень серьезными доводами в пользу общения с компьютером именно на естественном языке. Пока исследования носили чисто экспериментальный характер, эти доводы были достаточны. Однако в наше время, для которого характерна практическая переориентация работ, возникают новые проблемы, ранее остававшиеся в тени.

Часть из них: необходимость отчуждения системы от разработчика, надежность и устойчивость ее функционирования, эффективность реализации, наличие средств сопровождения – возникает и при создании традиционного программного обеспечения. Новые моменты связаны с использованием для общения с машиной именно естественного языка.

Среди проблем, особо актуальных на нынешнем этапе исследований и разработок, укажем:

- тщательный анализ вопроса целесообразности использования ЕЯ в человеко-машинном общении;
- поиск ситуаций, в которых общение с машиной на ЕЯ оправдано технологически и эргономически;
- выявление обстоятельств, учет которых обеспечивает человеку комфортные, естественные условия общения с компьютером;
- анализ пригодности использовавшихся ранее подходов и методов в изменившихся (практическая переориентация) условиях.

Перед автором некоторого искусственного языка общения с машиной (например, языка программирования), конечно же, не стоит вопрос о целесообразности использования созданного языка по прямому назначению. При оценке такого языка речь может идти о выразительных средствах, эффективной реализуемости, легкости усвоения и т.п. Отдельные неудачные решения могут быть изменены в ходе доработки (и отражены в разного рода пересмотренных сообщениях и др.). Объективация языка заключается в создании стандартов, трансляторов, формировании круга пользователей.

Естественный же язык изначально дан разработчикам компьютерных систем извне, он объективирован (и активно используется в речевой практике) в большой социальной группе носителей данного языка, которые привыкли к вполне определенным, человеческим условиям общения (в том числе, рассмотренным в начале данной главы). Если эти условия (человеческий фактор) будут игнорироваться, язык общения, возможно, сохранив внешнее сходство с тем или иным ЕЯ, потеряет главное – естественность. А учет этих условий требует от разработчиков систем очень серьезных дополнительных усилий, поскольку предполагает воссоздание (моделирование) нетривиальных человеческих механизмов работы с языком, наделение системы – как «собеседника» пользователя – основными чертами (на уровне информационных процессов) собеседника-человека.

Поэтому при создании компьютерных систем практической ориентации следует тщательно проанализировать, оправданы ли интеллектуальные и материальные затраты (весьма значительные, в нынешних условиях отсутствия в нашей стране рынка готового программно-информационного обеспечения) на их разработку, экономична ли (с учетом ресурсоемкости) их эксплуатация.

Серьезная практическая задача обеспечения общения с ЭВМ на естественном языке требует серьезного и практичного подхода. В каждой конкретной ситуации необходимо учитывать основательность доводов в пользу общения с системой именно на естественном языке, помнить о реально предоставляемых пользователю удобствах (в частности, об утомительности клавиатурного ввода, о возможностях – пока весьма скромных – технических средств обеспечения общения: устройства распознавания и синтеза звучащей речи, читающие автоматы).

Стремление разработчика или заказчика не отстать от моды, создать «высокоинтеллектуальную» информационную систему, оснащенную средствами естественноязычного интерфейса, не является достаточно веским основанием, а дилетантский подход (в этой новой и чрезвычайно сложной области особенно) не только не приводит к успеху, но и дискредитирует саму идею общения с ЭВМ на естественном языке.

Рассмотрим особенности естественного языка, осложняющие его использование для общения с ЭВМ (и, разумеется, задачу автоматической обработки текстов):

- ЕЯ – **большая система** (как термин),
- ЕЯ – **иерархическая система**,
- ЕЯ – **открытая система**,
- Связи элементов языка обычно **неоднозначны**,
- Отдельные носители языка используют **индивидуальные модели языка**,
- Использование языка сопровождается **речевыми ошибками**,
- **Описания языка** (построенные специалистами – лингвистами) **не полны и не точны**.

Задачу обеспечения **естественного общения** человека с машиной можно принять без каких бы то ни было оговорок. Однако ниоткуда не следует, что наиболее удобным и естественным для пользователя (и целесообразным, с точки зрения разработчика) средством такого общения будет естественный язык. Пререкания с «непонятливым» компьютером, ориентированным на ведение диалога с человеком на так называемом ограниченном естественном языке, могут потребовать более значительных усилий, чем изучение искусственного формального языка общения.

Нас интересуют ситуации, в которых необходимость использования естественного языка диктуется глубинными внутренними причинами:

- характер поручаемых системе заданий, а следовательно, и адресуемых ей сообщений таков (носит предварительный, неформальный характер), что описать их на каком-либо формализованном языке крайне трудно;

- общение пользователя с машиной происходит эпизодически и/или в очень широкой сфере (изучение специального языка общения нецелесообразно, или же он становится необъятным).

Часто в таких случаях альтернативой использованию естественного языка может служить хорошо разработанная схема «выспрашивания» нужной системе информации с помощью традиционных средств организации дружественного интерфейса (меню, опережающий ввод и др.).

Весьма интересен феномен появления **гибридных знаковых систем**.

Мы знаем, что конструкции естественного языка (слова, словосочетания, предложения, тексты) давно и широко используются в человеко-машинном интерфейсе.

При формулировке общих требований к диалогу часто отмечается, что диалог должен вестись на родном языке пользователя (или на другом понятном и привычном ему языке).

Вспомним:

- служебные слова в языках программирования, командных и других языках;
- названия позиций меню, кнопок и т.п.;
- средства поддержки пользователя:
  - сообщения об ошибках,
  - справочная информация,
  - внешняя документация.

Можно ли в подобных ситуациях говорить об общении с ЭВМ на естественном языке?

Вряд ли. Средства общения с традиционными компьютерными системами требуют предварительной экспликации, формализации той ситуации, о которой идет речь, привлечения знаний об используемых в системе способах структуризации знаний. Формулируя же адресуемые машине сообщения на естественном языке, человек может не знать: каким набором семантических единиц располагает компьютерная система, в какие формальные структуры будет отображаться содержание сообщения; какие компоненты описываемой ситуации являются в настоящий момент значимыми, а какие – второстепенными.

Более того, оказавшись в абсолютно новых условиях, человек, как правило, не сможет воспользоваться штатными языками общения с компьютером. Эти обстоятельства могут оказаться решающими при выборе языка общения с машиной как для профессионала (который на очередном этапе решения задачи не располагает пока адекватной формальной моделью и вынужден по этой причине отказаться от привычного для него искусственного языка), так и для конечного пользователя (для которого содержательный уровень общения является единственно доступным).

Лингвистическое и алгоритмическое обеспечение общения с компьютером на естественном языке

Для того, чтобы привлекательность диалога с компьютером на естественном языке не просто декларировалась, а стала реальностью, необходимо не только тщательно проанализировать условия и сценарии общения, но и:

- построить формальную модель естественного языка (учитывающую все его ключевые особенности),
- описать нетривиальное подмножество естественного в рамках этой модели,
- разработать и реализовать алгоритмы анализа и синтеза текстов.

Один из главных путей развития функциональных возможностей систем общения и повышения качества их работы – создание и внедрение более полных и точных моделей естественных языков, более совершенных алгоритмов анализа и синтеза текста.

Лингвистические банки данных/знаний

Под ***лингвистическими банками данных/знаний*** (ЛБД/ЛБЗ)[1] понимаются представленные в электронной форме языковые источники (корпусы текстов) и лингвистические описания.

Отметим, что в наше время, в ситуации, когда надежность работы систем оптического распознавания близка (на хороших по качеству печатных текстах) к 100%, в электронную форму легко переводимы и традиционные источники информации о языке.

Говоря о системах распознавания, мы, разумеется, имеем в виду не системы сканирования или фотографирования печатных или рукописных текстов (результатами работы которых являются графические файлы в форматах DJVU, JPEG (он же JPG) и др.), а средства получения результатов распознавания в текстовых форматах.

Поскольку такие средства существуют и широко используются на практике, можно считать, что в ЛБД можно перевести любые полиграфические источники: тексты на том или ином естественном языке, словари, справочники, книги по лингвистике.

Спектр ЛБД достаточно широк: это как необработанные («сырые») корпусы текстов, так и тексты с некоторыми добавлениями, например грамматическими характеристиками слов, стилистическими пометами (разговорное, специальное и т.п.), или описаниями синтаксической структуры предложений (соответствующие корпусы текстов называют ***размеченными***). Сюда также входят разнообразные компьютерные словари: частотные, грамматические, словоформ, тезаурусы, словари словосочетаний и моделей управления, своды грамматических правил и т.п.

Различаться может и назначение лингвистических банков/баз данных. Часть из них предназначена для автоматизации деятельности лингвистов и разработчиков прикладных систем, часть – для непосредственного использования в системах

[1] Оба термина (ЛБД и ЛБЗ) обычно используются в научной литературе как синонимы, хотя отдельные авторы с этим не согласны.

обработки текста и речи: автокорректорах, системах распознавания текста и речи, информационно-поисковых системах.

Подробнее о ЛБЗ мы поговорим позже.

Практически значимые сферы применения систем автоматической обработки текстов

Системы ***автоматической обработки текста*** (***АОТ-системы***) по выполняемым функциям (входной и выходной информации) можно классифицировать следующим образом:

	Язык входного текста	Язык выходного текста
1	Естественный-1	Естественный-2
2	Искусственный	Естественный
3	Естественный	Искусственный / Естественный
4	Естественный	Естественный + { Искусственный}

К системам первого типа относятся программы машинного перевода, получающие текст на некотором естественном языке и перерабатывающие его в текст на другом естественном языке. Второй тип – системы генерации (синтеза) текстов по некоторому формальному описанию. Системы третьего типа, наоборот, перерабатывают текст на естественном языке в текст на искусственном (индексирование, извлечение смыслового содержания) или в другой текст на естественном языке (реферирование). К последнему классу отнесем программы, занимающиеся проверкой текста, написанного на естественном языке. Они в результате своей работы либо исправляют входной текст автоматически, либо формируют некоторый протокол замечаний.

Естественный язык - сложная, многоплановая система, с множеством правил, внутренних связей, имеющая отношение ко всем аспектам деятельности человека. Точность и правильность работы программ определяется глубиной анализа. Достаточно глубокий анализ пока достигается только для определенных узких предметных областей (из-за специфичности подъязыка такой области: в каждой области свои термины, специфические семантические отношения и т.п.).

Для создания систем, работающих со всем естественным языком без потери глубины анализа, в настоящий момент не хватает либо технических возможностей (быстродействия, памяти), либо теоретической базы (например, пока нет даже единой схемы достаточно полного, глубокого и непротиворечивого описания семантики естественного языка). Однако в коммерческих системах, ввиду того, что предназначаются они для большого количества пользователей, разных предметных областей, принята концепция поверхностного анализа, к тому же и производится такой анализ значительно быстрее. Дальнейшее продвижение вперед, использование естественного языка в практических областях невозможно без оснащения этих систем обширными и глубокими (с точки зрения охвата различных явлений языка) описаниями и моделями, созданными лингвистами-профессионалами.

Эта тенденция прогнозируется многими исследователями и прослеживается на примере развития АОТ-систем, уже в наши дни представляющих коммерческий интерес и использующихся при решении следующих прикладных задач:

1. **Machine Translation and Translation Aids** - машинный перевод;
2. **Text Generation** - генерация текста;
3. **Localization and Internationalization** - локализация и интернационализация;
4. **Controlled Language** - работа на ограниченном языке;

5. **Word Processing and Spelling Correction** - создание текстовых документов (ввод, редактирование, исправление ошибок)

6. **Information Retrieval** - информационный поиск и связанные с ним задачи.

Отметим, что это деление несколько условное, и в реальных системах часто встречается объединение функций. Так, для машинного перевода требуется генерация текста, а при исправлении ошибок приходится заниматься поиском вариантов словоформы и т.д.

Машинный перевод

Исторически ***машинный перевод*** является первой попыткой использования компьютеров для решения невычислительных задач (знаменитый Джорджтаунский эксперимент в США в 1954 г.; работы по машинному переводу в СССР, начавшиеся в 1954 г.). Развитие электронной техники, рост объема памяти и производительности компьютеров создавали иллюзию быстрого решения этой задачи. Идея захватила воображение ученых и администраторов. Практическая цель была простой: загрузить в память компьютера максимально возможный словарь и с его помощью из иноязычных текстов получать текст на родном языке в удобочитаемом виде. Однако первоначальная эйфория по поводу того, что столь трудоемкую работу можно поручить ЭВМ, сменилась разочарованием в связи с абсолютной непригодностью получаемых текстов. Приведем в качестве примера результаты работы одной из современных коммерческих систем перевода. Предложим ей перевести народное английское стихотворение, известное нам в переводе «Робин-Бобин» (текст этот очень простой, московские дети изучают его в начальной школе):

Robin, Robin, what a man!	*Малиновка, Малиновка, какой человек!*
He eats as much as no one can.	*Он ест насколько никто не может.*
He ate a lot of fish, he ate a lot of meat.	*Он съел много рыб, он съел много мяс.*
He ate a lot of ice-cream and a sweet.	*Он съел много ледяных-сливки и сладкий.*
He ate a lot of porridge and ten eggs	*Он съел много каша и десять яйцо*
And all the cookies Mother had.	*И вся Мать повары имела.*
He drank a lot of juice, he ate a cake	*Он пил много соков, он съел торт*
Then said: «I have a stomach-ache»	*Затем сказал: «У меня есть желудок-боль»*

Сравним с художественным переводом К.Чуковского:

Робин Бобин Барабек
Скушал сорок человек.
И корову, и быка,
И кривого мясника,
И телегу, и дугу,
И метлу, и кочергу.
Скушал церковь, скушал дом,
И кузницу с кузнецом,
А потом и говорит:
- У меня живот болит!

Следующий пример показывает неустойчивость системы машинного перевода при обработке неоднозначностей. Два предложения по отдельности *Flyer flies*" и *Flyers fly*" переводятся так: *Летчик летает* и *Летчики летают*, если же из тех же словосочетаний составить одно предложение *Flyer flies and flyers fly* получаем *Летчик летает и муха летчиков.*

Конечно, системы, настроенные на определенную предметную область, дают гораздо более приемлемые результаты. Однако в этом случае системы перевода получаются очень узко ориентированными, и попытка использовать их даже в смежных предметных областях дает совершенно непредсказуемые результаты. Подобные эксперименты даже распространены среди любителей пошутить: инструкция по эксплуатации манипулятора-мыши, переведенная с английского языка на русский системой автоматического перевода, использующей специализированный медицинский словарь, превращается в описание всевозможных издевательств над несчастным маленьким грызуном.

Возникают эти проблемы из-за принципиально разных подходов к переводу человека и машины. Квалифицированный переводчик понимает смысл текста и **пересказывает** его на другом языке словами и стилем, максимально близкими к оригиналу. Для компьютера этот путь выливается в решение двух задач: 1) перевод текста в некоторое внутреннее семантическое представление и 2) генерация по этому представлению текста на другом языке. Поскольку не только не решена сама по себе ни одна из этих задач, а нет даже общепринятой концепции семантического представления текстов, при автоматическом переводе приходится фактически делать "подстрочник", заменяя по отдельности слова одного языка на слова другого и пытаясь после этого придать получившемуся предложению некоторую синтаксическую согласованность. Смысл при этом может быть искажен или безвозвратно утерян.

Более реалистичными являются попытки создать системы ***автоматизированного перевода*** - программы, которые не берут на себя полностью весь перевод, а лишь помогают человеку-переводчику справиться с некоторыми трудностями (Computer Aided Translation). Одним из примеров таких систем является Eurolang Optimizer. Его можно рассматривать как нечто переходное между компьютерным словарем и программой-переводчиком, как некий набор предметно-ориентированных глоссариев, снабженный интерфейсом для удобства переводчика: предлагается несколько вариантов перевода, выделенные разными цветами в зависимости от условий применимости; переводчик может с помощью меню определенным образом настраивать словари для более быстрого и правильного выбора нужного эквивалента.

Подобные программные средства могут помочь в решении проблем, связанных с терминологией и вообще со знаниями переводчика о предметной области: одни и те же слова могут по-разному переводиться в зависимости от того, о каком предмете идет речь.

Автоматически может быть решена проблема согласованности. Понятно, что согласованность важна в рамках одного документа: один и тот же термин, даже если его без потери смысла можно перевести несколькими словосочетаниями, должен переводиться одинаково на протяжении всего документа. Однако еще более важной

является согласованность в широком смысле - разработка и применение единой концепции интерпретации одного и того же термина на разных языках (скажем, американский разработчик программного обеспечения может быть недоволен, что термин *dialog box* переводится на итальянский как *finestra* (окно) и как *boite* (коробка, ящик) на французский). Ошибки, возникающие вследствие нарушения согласованности, являются серьезной проблемой, так как, имея только текст-результат перевода, уже невозможно установить, какие термины в оригинале были одинаковыми, а теперь переведены по-разному (в отличие от орфографических ошибок, которые исправить никогда не поздно).

В последнее время также появляются автоматизированные системы «доперевода» или «перевода изменений». Их возникновение связано с тем, что большинство технических текстов (описания, инструкции) не являются целиком новыми (как и явления, продукты, механизмы и т.п., ими описываемые), а содержат в себе лишь некоторые изменения, связанные, например, с усовершенствованием конструкции. Система «доперевода» извлекает из памяти знакомые предложения, а новые куски предлагает переводчику. Заметим, что такой человеко-машинный способ генерации новых текстов также помогает согласованности в стиле и терминологии при переходе от одной версии к другой.

Развитием систем подобного вида можно считать канадскую (Канада – двуязычная страна, постоянно сталкивающаяся с проблемой перевода на государственном уровне) систему генерации прогнозов погоды **Forecast Generator** (**FOG**). Можно считать, что в ней перевод полностью заменен генерацией текстов. В памяти системы хранится 20 миллионов слов и словосочетаний, связанных с прогнозами погоды, что позволяет генерировать как английский, так и французский вариант непосредственно из базы данных. Конечно, успешная работа этой системы в значительной мере объясняется ограниченной природой текстов: сообщения о погоде являются классическим примером подъязыка. Ограниченность словаря, грамматики и семантики дает возможность достичь отличных результатов сравнительно простыми методами.

Генерация текста

С необходимостью генерации хотя бы простейших фраз разработчики практических систем столкнулись еще на заре их создания. Даже в столь примитивно организованной (в плане дружественности пользовательского интерфейса) среде, как DOS, при попытке сгенерировать стандартное сообщение о количестве скопированных файлов мы сталкиваемся с проблемой построения фразы: в зависимости от этого количества необходимо использовать разные слова (в английской версии *file* в случае одного файла и *files*, если больше; в русской - и того хуже: могут встретиться варианты *файл*, *файла* и *файлов*, причем правила, в каком случае какой из них использовать, достаточно сложны).

По степени сложности и выразительности существующие методы генерации сообщений принято подразделять на 4 класса (часто используются комбинации методов). Рассмотрим их на примере генерации сообщений о копировании файлов.

1) Canned-based methods

Неизменяющийся шаблон - просто печать строки символов без каких-либо изменений.

Для генерации сообщений создаются таблицы шаблонов, которые будут выдаваться в зависимости от ситуации. В нашем варианте при копировании одного файла будет напечатана первая строка таблицы:

1 file copied,

а в случае, например, трех - третья:

3 files copied

2) Template-based methods

Изменяющийся шаблон - бесконтекстная вставка слов в образец-строку (именно этот метод используется в MS-DOS):

Шаблон: **<Число> file(s) copied**

может быть использован для генерации сообщений:

0 file(s) copied,

1 file(s) copied,

2 file(s) copied

3) Phrase-based methods

Контекстная вставка.

В зависимости от вида сообщения (контекста) шаблон может быть несколько изменен. Скажем, система может распознавать, с каким окончанием писать слово *file* в зависимости от их количества.

Шаблон: **<Число> <Определение> <file/files при =1, >1><Глагол: время - прош.>**

может использоваться для генерации сообщений:

1 file copied,

2 marked files copied,

2 marked files deleted

4) Feature-based methods

Синтез сообщения на основе набора свойств (грамматических признаков).

Это наиболее сложный метод, он требует привлечения обширных лингвистических знаний, но, в то же время, он и наиболее привлекателен. Предложение определяется набором характеристик составляющих его слов (например, наличие/отсутствие отрицания, настоящее/прошедшее время) и правилами их сочетаемости.

Шаблон: **<Число><Определение><file/files при =1, >1><Глагол: время - любое>**

позволяет генерировать сообщения:

1 file should be copied,

1 file was copied,

2 marked files were copied

Понятно, что генерация логически связных, целостных текстов является гораздо более сложной задачей: к правилам построения предложений добавляются правила их сочетаемости, правила развития сюжета, соблюдения стиля и т.п. Ввиду невозможности их полной формализации задачу генерации полноценных художественных текстов можно считать на настоящий момент неразрешимой. Однако для некоторых специализированных технических текстов эти правила строго оговорены некоторыми стандартами, немногочисленны и поэтому поддаются формализации. Примером таких текстов могут служить различные инструкции, техническая документация, тем более задача ее автоматической генерации давно назрела.

На Западе уже давно разработка документации превратилась в особую подотрасль разработки любых достаточно сложных систем (в том числе программного обеспечения). Сопроводительная техническая документация весьма разнообразна: руководство пользователя, руководство для менеджера (администратора) системы, руководство по монтажу (инсталляции) и первичному запуску, руководство по эксплуатации, руководство по интегрированию системы с другими устройствами (программами), проектные материалы и т.д. Однако часто пользователь не получает своевременно и в полном объеме необходимый ему материал, соответствующий используемой им версии системы. Это можно объяснить двумя причинами. Во-первых (субъективная причина), подготовка документации - это дополнительная работа, требующая дополнительного времени и дополнительных навыков (разработчику трудно изложить требуемое на понятном рядовому пользователю языке, остальным же надо сначала детально изучить систему). Во-вторых (объективная причина), документация устаревает по ходу модернизации системы.

Поиски решения этих проблем привели в свое время к появлению новой профессии «технического писателя». Однако понятно, что привлечение дополнительных работников ведет к удорожанию продукта. Поэтому в последние годы появились практические системы, осуществляющие помощь в разработке документации, вплоть до ее автоматической генерации. Форма и содержание документации часто выбирается не столько из соображений удобства и полезности для пользователя, сколько из соображений простоты ее создания.

Документация, как правило, содержит графическую и текстовую части. Графическую часть проще сформировать, однако без текстовой не обойтись: в ней описывается семантика продукта (назначение, технические данные, ограничения, детализация работы в разных режимах). Очевидно, что качественная система должна генерировать текст, правильный с точки зрения грамматики и синтаксиса естественного языка. Поскольку предметная область точно определена, а техническая документация составляется по определенным строго заданным правилам, степень формализации в постановке данной задачи существенно выше, чем в задаче машинного перевода, что позволяет надеяться на более высокие результаты.

Локализация и интернационализация

Для того чтобы иметь успех на международном рынке, программные продукты должны быть локализованы, т.е. приспособлены к культурным и языковым нормам потенциальных покупателей.

Для многих программных приложений локализация может быть сравнительно простой, когда основная программа (алгоритм) изменяется незначительно. Конечно, опции меню, сообщения об ошибках, экранные подсказки и другие текстовые строки, вставленные в программу, должны переводиться, но это не создает особых проблем, если при разработке приложения была предусмотрена возможность локализации. Для решения этой задачи программный код и текст должны быть разделены. По установленному стандарту текстовые строки оформляются в отдельном файле, вызываемом из программы. Таким способом текстовые строки можно переводить, не затрагивая исходный код.

Подобные принципы облегчения локализации возможны не для всех приложений. Системы, в которых естественный язык используется не только для формирования сообщений на экране, но и является предметом деятельности самой системы (например, программы-автокорректоры), поддаются локализации с большим трудом. Здесь могут потребоваться большие специализированные словари и полная переработка алгоритмов. Часто эта задача настолько сложна, что разработчик ею заниматься не может, и проблема локализации приложений является заботой пользователя-носителя языка.

В идеале для нашего многоязычного мира программные средства должны быть интернациональными; пользователь, купив версию программы для некоторого языка, не должен покупать другую версию для другого. Назрела необходимость иметь программные средства, позволяющие автоматически настраивать приложение на заданный язык. Пока мы еще далеки от этой цели, но работы в этой области ведутся с большой интенсивностью, особенно в Европе, где в связи с образованием Европейского Союза возникает необходимость вести дела и документацию на всех официальных и некотором количестве неофициальных языков.

Работа на ограниченном языке

Одним из способов разрешения проблем, связанных с обработкой естественного языка, является упрощение и некоторая формализация самих текстов: использование ограниченного языка (подмножества языка). Под ограниченным понимается упрощенный язык, использующий ограниченный словарь, грамматику, строго определенные несложные синтаксические конструкции. Обычно в нем запрещаются длинные предложения, длинные цепочки существительных (типа *решение проблемы разработки систем перевода на базе представления текста в виде последовательности предложений*...), не используются пассивные и негативные конструкции, вводятся строгие правила использования терминов. Тексты должны соответствовать одному из стандартных стилей или даже быть составлены по определенному шаблону, принятому в для документов подобного рода.

Эти правила не являются современным изобретением: именно их обычно применяют при написании технической документации. Достаточно «древним» примером ограниченного языка является «Бэйсик Инглиш», введенный англичанами для общения с туземным населением в колониях. Неожиданно он оказался полезен

и для общения самих туземцев друг с другом: колонизация ввела в их быт множество предметов и понятий, просто не имеющих названий в их родных языках. Забавно, что через много лет при «колонизации» Европы и всего мира англоязычными техническими средствами используются практически те же методы. Например, все специалисты в области компьютерной техники пользуются английскими терминами (*файл*, *принтер* и т.д.), не пытаясь подыскать эквивалент на родном языке, и мы по-русски говорим *word для windows*, а не *слово для окон*.

Применение ограниченного языка делает документ более понятным, удобным для восприятия, он становится легче для переводчиков, поскольку дает меньше возможностей для неоднозначного толкования: такой документ легче составить автору, не являющемуся носителем языка документа. Правительства, особенно в Европе, начинают вводить стандарты на подготовку документации, нормы, по которым требуется использование ограниченных языков, особенно в международной торговле. В связи с этим возникает потребность автоматизации проверки соответствия текста правилам ограниченного языка, создания систем «перевода» с естественного языка на ограниченный.

Boeing, Caterpillar и несколько других компаний призвали вести всю документацию только на ограниченном языке. Ими разработана система Boeing Simplified English Checker для проверки соответствия текстов различным промышленным стандартам и государственным нормам. На ее базе создается программа Clearcheck, не только контролирующая правильность текста на ограниченном языке, но и исправляющая ошибки.

Некоторые разработчики прогнозируют создание систем с использованием ограниченных языков, в которых полный и корректный перевод документации будет производиться без вмешательства человека.

Создание текстовых документов (ввод, редактирование, исправление ошибок)

Нет необходимости говорить о многообразии систем для подготовки текстовых документов: текстовых редакторов, издательских систем и т.п. Они прочно вошли в нашу жизнь, без них не может обойтись ни один пользователь и ни одна область деятельности. Более того, создание текстовых документов – одна из основных сфер применения персональных компьютеров. Использование текстовых редакторов обусловлено не только тем, что они облегчают работу, но и тем, что в последнее время во многих сферах деятельности введены стандарты на подготовку текстов, основанные на применении определенных редакторов.

В отличие от машинного перевода разработка систем редактирования текстов еще на заре своего развития, в 60-е годы, считалась коммерчески перспективной прикладной областью. В настоящее время рынок перенасыщен подобными системами; среди их создателей существует жесткая конкуренция, поэтому при введении одним из поставщиков каких-либо новых возможностей (например, проверка стиля) остальные вынуждены вводить в свои системы нечто подобное. Одним из первых массовых нововведений стало включение в состав текстового редактора программ проверки правописания и внесения необходимых исправлений – ***автокорректоров***. Чтобы придать своему продукту новые коммерчески перспективные свойства, создатели вынуждены все активнее использовать

лингвистические знания, применять методы морфологического и синтаксического анализа. На очереди - создание систем, выполняющих функции научного редактора, т.е. осуществляющих литературную и научную правку текстов, способных производить сложное автоматизированное редактирование текстов.

Проверка текста в таких системах может вестись в режиме off-line – когда формируется протокол замечаний по тексту, либо в режиме on-line – когда исправление ошибок ведется по мере их обнаружения (возможно, после получения соответствующего подтверждения от пользователя). При обнаружении ошибки система может предложить вариант ее исправления (при наличии нескольких вариантов – их упорядоченный список). Замечания по тексту также могут носить различный характер. Они могут быть локальными (указывается фрагмент текста с ошибкой) и глобальными (выдается диагностическое сообщение, касающееся всего текста, например: «текст труден для восприятия»).

Поиск информации и связанные с ним задачи

Не вызывает сомнений необходимость автоматизации поиска заданных текстовых фрагментов в текстах на естественном языке. Однако часто даже при поиске информации другого рода (например, аудио- и видео-) работа на самом деле ведется с описаниями на естественном языке (например, для организации поиска фотографий необходимо снабдить каждую из них набором словесных характеристик типа «*портрет, профиль, полный рост, женщина*», «*пейзаж, лес, осень*» и т.п.).

В последних разработках классических систем поиска текста основное внимание уделяется дополнению их разнообразными средствами текстовой обработки, что приводит к расширению возможностей и облегчению работы для пользователя-непрофессионала.

Применение компьютеров не только ускоряет создание и обработку документов, но и чрезвычайно стимулирует рост их количества и объема. Очень многие пользователи регулярно сталкиваются с необходимостью быстро просматривать большой объем документов и выбирать из них действительно нужные. Эта задача возникает при работе с текстовыми базами данных, с электронной почтой, при поиске в Интернете. Сократить количество просматриваемых документов могут помочь системы автоматического ***рубрицировая***. Поток входных документов эти системы распределяют по небольшому количеству классов (тематических рубрик). При этом могут учитываться как чисто внешние показатели документов (объем, расширение имени соответствующего файла и т.п.), так и их содержательные характеристики (название, фамилия автора, ключевые слова), которые могут позволить отнести текст к той или иной рубрике.

Часто в крупных организациях, особенно государственных, правила делопроизводства предписывают сопровождать каждый документ кратким описанием или набором ключевых слов. Во всех указанных случаях была бы весьма полезна возможность автоматически составлять сжатые описания содержания документов – рефераты.

К сожалению, автоматические методы не настолько совершенны, чтобы создать полноценный реферат путем генерации предложений текста. Однако уже сейчас

возможно ***автоматическое реферирование*** – составление более или менее информативных и связных рефератов заданного объема (***квазирефератов***) – путем выбора информативных предложений из исходного текста, а также выделение достаточно представительного списка ключевых слов.

В качестве ключевых слов система может выбирать слова, наиболее часто встречающиеся в тексте (и являющиеся при этом информативными, т.е. не предлоги, союзы и проч.), либо использовать для отбора какие-либо синтактико-семантические признаки (из фрагмента: *Определение. Интегралом ... называется ...* можно заключить, что *интеграл* – ключевое слово).

При реферировании из текста отбираются предложения, в наибольшей степени характеризующие его содержание. Например, предложения, содержащие ключевые слова (чем больше, тем лучше), либо отобранные по некоторым особым признакам. Размер реферата (коэффициент сжатия) или количество ключевых слов задается пользователем. Результатом работы такой системы может являться некоторый новый текстовый документ (реферат или набор ключевых слов) или же данный документ, в котором ключевые слова или наиболее информативные предложения выделены по тексту.

Возвращаемся к вопросам роли и места естественного языка в человеко-машинном общении.

Как мы уже отмечали: ***Серьезная практическая задача обеспечения общения с ЭВМ на естественном языке требует серьезного и практичного подхода.*** В каждой конкретной ситуации необходимо учитывать основательность доводов в пользу общения с системой именно на естественном языке, помнить о реально предоставляемых пользователю удобствах (в частности, об утомительности клавиатурного ввода, о возможностях – пока весьма скромных – технических средств обеспечения общения: устройства распознавания и синтеза звучащей речи, читающие автоматы).

Стремление разработчика или заказчика не отстать от моды, создать «высокоинтеллектуальную» информационную систему, оснащенную средствами естественноязычного интерфейса, не является достаточно веским основанием, а дилетантский подход (в этой новой и чрезвычайно сложной области особенно) не только не приводит к успеху, но и дискредитирует саму идею общения с ЭВМ на естественном языке.

Лингвистическое и алгоритмическое обеспечение общения с компьютером на естественном языке.

Для того, чтобы привлекательность диалога с компьютером на естественном языке

не просто декларировалась, а стала реальностью, необходимо не только тщательно

проанализировать условия и сценарии общения, но и:

- построить формальную модель естественного языка (учитывающую все его ключевые особенности),
- описать нетривиальное подмножество естественного в рамках этой модели,
- разработать и реализовать алгоритмы анализа и синтеза текстов.

Один из главных путей развития функциональных возможностей систем общения и повышения качества их работы – создание и внедрение более полных и точных моделей естественных языков, более совершенных алгоритмов анализа и синтеза текста.

Лингвистические банки данных и базы знаний.

Под ***лингвистическими банками данных*** (ЛБД) понимаются представленные в электронной форме языковые источники (корпусы текстов) и лингвистические описания.

Иногда в достаточно близком значении используется термин, взятый из работ по искусственному интеллекту, – ***Лингвистические базы знаний***.

Отметим, что в наше время, в ситуации, когда надежность работы систем оптического распознавания близка (на хороших по качеству печатных текстах) к 100%, в электронную форму легко переводимы и традиционные источники информации о языке. Поэтому можно считать, что в ЛБД можно перевести любые полиграфические

источники: тексты на том или ином естественном языке, словари, справочники, книги по лингвистике.

Спектр ЛБД достаточно широк: это как необработанные («сырые») корпусы текстов, так и тексты с некоторыми добавлениями, например грамматическими характеристиками слов, стилистическими пометами (разговорное, специальное и т.п.), или описаниями синтаксической структуры предложений (соответствующие корпусы текстов называют ***размеченными***). Сюда также входят разнообразные компьютерные словари: частотные, грамматические, словоформ, тезаурусы, словари словосочетаний и моделей управления, своды грамматических правил и т.п.

Различаться может и назначение лингвистических банков данных. Часть ЛБД предназначена для автоматизации деятельности лингвистов и разработчиков прикладных систем, часть – для непосредственного использования в системах обработки текста и речи: автокорректорах, системах распознавания текста и речи, информационно-поисковых системах.

Два типа технологий формирования ЛБД:

- Формирование "вручную" экспертами (филологи, специалисты в области применения компьютерной системы); здесь нужны инструментальные средства поддержки их работы.
- Формирование в автоматизированном режиме с использованием методов машинного обучения; желательно участие экспертов для анализа и коррекции результатов, полученных в автоматическом режиме.

Об автоматизированном формировании лингвистических баз знаний

В зависимости от решаемых задач ЛБЗ могут включать различные словари[2], отличающиеся друг от друга глубиной проникновения в структуру описываемого языка и характером содержащейся информации.

Можно выделить несколько уровней глубины проникновения в структуру языка, отражаемых в компонентах ЛБЗ.

Лексический уровень	Словари словоформ
	Частотные словари словоформ
	Словари основ и неизменяемых слов
	Словари лексических n-грамм
Синтаксический уровень	Грамматические словари
	Частотно-грамматические словари
	Словари грамматических n-грамм
	Словарь синтаксем и синтаксические правила
	Словари моделей управления
Семантический уровень	Толковые словари
	Словари семантических моделей управления

[2] Под термином “***словарь***” мы понимаем (если явно не оговорено иное) компьютерный словарь, описывающий явления любого языкового уровня.

	Тезаурусы

Компоненты лексического уровня содержат информацию об особенностях отдельных слов языка, например, об их морфологических свойствах, частотности, лексической сочетаемости. На синтаксическом уровне представлены знания о правилах связывания лексем в более сложные синтаксические конструкции – словосочетания, предложения, информация о синтаксических свойствах лексем (например каждой лексеме может быть приписан некоторый синтаксический класс и указаны правила сочетания представителей этих классов в предложении). Семантический уровень содержит информацию о связях понятий предметной области.

Состав ЛБЗ

Следует отметить, что распределение словарных компонентов по уровням является условным. Так, например, лексические n-граммы несут информацию о статистических характеристиках их сочетаемости, отражающую явления и закономерности синтаксического уровня.

Мы рассмотрим некоторые аспекты проблемы автоматизированного формирования лингвистических баз знаний на примере таких их компонентов как информационно-поисковый тезаурус и словарь лексических биграмм.

Отметим, что в последние годы для формирования других компонентов ЛБЗ широко применяются (иногда и весьма успешно) методы машинного обучения.

Можно указать следующие особенности ЛБЗ современных систем, предназначенных для работы в предметных областях (ПО) реальных масштабов сложности:

- большой объем лингвистической информации, необходимый для покрытия явлений подъязыка ПО;
- высокие требования к точности обработки естественно-языкового материала;
- постоянное обновление ЛБЗ, необходимое для поддержания актуальности хранящихся данных.

Примером задачи, требующей анализа большого потока текстов из широкого круга ПО, предъявляющей высокие требования к качеству обработки лингвистической информации и актуальности ЛБЗ может служить проблема поддержки работы аналитика, получающего большой поток текстовых документов и исследующего всю совокупность документов в их взаимосвязи.

Из требований, вытекающих из первого и третьего пунктов, следует необходимость автоматизации формирования и сопровождения ЛБЗ, поскольку создание и, тем более, поддержание ЛБЗ в актуальном состоянии требуют огромных усилий. Второе требование накладывает условия на качество представленной в ЛБЗ информации и, соответственно, на методы, которые применяются для формирования ЛБЗ. Так, например, в системах распознавания устной речи статистические критерии выбора наиболее подходящего варианта распознавания обеспечивают до 90-95% правильно распознанных слов. Для получения более точного результата необходимо использование более изощренных лингвистических средств, например таких, как локальный синтаксический анализ.

На текущем этапе развития компьютерной лингвистики формирование высококачественных ЛБЗ невозможно как без компьютерной поддержки, так и без участия человека. Важнейшей функцией систем автоматизации формирования ЛБЗ является помощь человеку в решении этой задачи – система должна взять на себя максимум рутинной работы.

Наиболее существенными из таких рутинных операций являются:

- сбор информации о лингвистических явлениях;
- анализ собранной информации с целью экспликации скрытых закономерностей между изучаемыми лингвистическими явлениями;
- предоставления результатов эксперту в удобной для него форме;
- организация коллективной работы экспертов по созданию и актуализации ЛБЗ.

Рассмотрим некоторые подходы к формированию информационно-поискового тезауруса. Их можно классифицировать по следующим признакам:

- используемые методы (статистические; лингвистические; использующие эвристики, отражающие особенности языка рассматриваемой ПО);
- анализируемые источники (произвольные тексты; специально подготовленные тексты, например аннотации или заголовки статей; текстовые источники, содержащие лингвистическую информацию, но ориентированные на ее использование человеком, а не на автоматическою обработку);
- роль экспертов в процессе формирования (оценка результатов и подбор параметров системы; построение базы эвристик; подготовка тренировочного корпуса текстов; коррекция полученной автоматически предварительной информации).

Наиболее простыми с точки зрения реализации и наименее точными являются чисто статистические методы, основанные на стохастической модели и результатах экспериментов на специально подобранном корпусе текстов.

Обычно рассматривается некоторое множество текстов и составляющих их терминов и на базе совокупности событий, описываемых частотными характеристиками вхождения терминов в текст (частотности терминов, их взаимной встречаемости) выделяются дескрипторы тезауруса и делаются заключения о наличии между ними некоторых тезаурусных связей. Такие подходы часто приводят к неплохим результатам, однако лишь при серьезных ограничениях на предметную область и характер использования тезауруса. Так, в одной из достаточно старых работ рассматривается одна из попыток статистического формирования информационно-поискового тезауруса, успех который был обусловлен:

- ограниченностью предметной области – один из разделов молекулярной биологии (такая узкая предметная область не балует разнообразием лингвистических явлений);
- серьезной предварительной работой по классификации искомых дескрипторов тезауруса (предварительно были выделены такие классы понятий, как имена ученых, названия генов и т.п.).

Важным направлением в автоматизированном формировании ЛБЗ является использование лингвистических источников не предназначенных для автоматической обработки. Неоднократно предпринимались попытки использовать обычные толковые словари (или Википедию) для формирования первого приближения к тезаурусу. При этом, например, рассматриваются специальные ссылки, которые обычно в словарях (англоязычных) выделяются особым образом, такие как: *see also*, *synonym with*, *see under* и т.п.

Устанавливается соответствие между подобными ссылками и тезаурусными отношениями, по словарю выделяются дескрипторы и связи между ними, устраняются противоречия (например, отсутствие в исходном словаре обратных ссылок между дескрипторами, соответствующих прямым). Затем результат предъявляется эксперту, который завершает формирование тезауруса, используя собственные знания.

Однако в толковых словарях могут использоваться для выявления потенциальных тезаурусных отношений между понятиями не только специальные ссылки между словарными статьями, довольно легко выявляемые по тексту словаря, но и информация заключенная в словесных формулировках толкований значений терминов. Толкование обычно раскрывает значение определяемого термина путем ссылок на значения некоторых других описываемых в словаре терминов. Причем для толкований характерны достаточно единообразные способы выражения связей между понятиями. Оказалось, что, исследуя текст толкования, можно выявить некоторые виды связей между терминами, например, родо-видовые, часть-целое.

По таким связям между парами понятий может быть построен граф, анализируя который можно получить предварительную классификацию терминологического состава словаря, которая может использоваться в качестве нулевого приближения к тезаурусу. Опыт работы нашей группы показал, что при разумном распределении обязанностей между компьютером и человеком можно подготовить за обозримое время приемлемый информационно-поисковый тезаурус.

Лингвистическая информация может быть успешно использована даже при коррекции таких традиционно чисто-статистических компонентов ЛБЗ, как словари лексических n-грамм. Словари n-грамм (обычно используются словари биграмм) часто применяются в задачах распознавания (речи или текстов по графическому представлению), где критерием эффективности является процент правильно распознанного языкового материала (например, слов).

Статистическая информация о встречаемости слов и их комбинаций может помочь выбрать наиболее вероятный омофон (даже при абсолютной надежности акустического и фонетического распознавания) или предпочтительный вариант словоформы (при неоднозначных результатах работы оптического распознавателя).

Для разрешения подобных неоднозначностей могут быть использованы и другие подходы. Например, синтаксический анализатор является более мощным средством улучшения результатов распознавания. Однако алгоритмы разрешения неоднозначности на основе информации из частотных словарей применяются очень широко, поскольку они проще и менее требовательны к вычислительным мощностям системы, чем алгоритмы синтаксического анализа.

Естественно возможность получения верного результата (например правильного выбора второго слова по распознанному первому слову биграммы) будет больше на

том корпусе текстов, на базе которого формировался данный словарь биграмм. Если множество текстов было достаточно представительно для некоторой предметной области, то полученная информация (биграммы и их веса) будет адекватна и большинству текстов этой ПО.

Обычно очень трудно подобрать тексты, покрывающие всю проблемную область, особенно для областей, в которых происходит интенсивное накопление знаний.

Выявляются новые свойства объектов исследования, новые объекты и, соответственно, появляются новые языковые единицы (например, терминологические сочетания), или изменяются старые (например, модели управления).

Слова и словоформы, встречающиеся в текстах ПО, в зависимости от частотности их встречаемости отображаются в словаре в набор биграмм с той или иной степенью адекватности отражающий особенности их сочетаемости в подъязыке предметной области. Лексемы, которые встречались в тексте достаточно часто, получают более адекватную информацию, реже встречающиеся – менее адекватную.

Так, например, некоторые допустимые в языке биграммы могут вообще не встретиться в тренировочном наборе текстов, над которым строился словарь биграмм. При работе системы распознавания это может привести к отбрасыванию некоторых правильных вариантов распознавания, что приводит к заметному ухудшению результатов, особенно в системах с высокими (более 90%) требованиями к надежности распознавания. Практика показала, что наличие возможной биграммы в словаре, даже с очень маленьким весом, дает возможность выбрать правильный вариант с гораздо большей вероятностью, чем при отсутствии такой биграммы.

В нашей исследовательской группе были разработаны методы пополнения словаря биграмм, позволяющие по существующим в словаре биграммам добавить возможные, но отсутствующие биграммы. Пусть, например, высокочастотное слово встречается в нескольких формах в биграммах словаря с некоторым другим словом, а в некоторых формах, являющихся допустимыми с точки зрения синтаксической сочетаемости этих слов, не встречается. В этом случае в словарь добавляются биграммы с недостающими формами и им присваивается небольшой вес в силу их редкой встречаемости. Этот метод был назван "пополнение парадигмы".

Другой метод – "трансформация" – применяется, когда биграмма представляет собой согласованное словосочетание или его часть. Если это словосочетание допускает перефразирование, из которого вытекает допустимость некоторой биграммы, которой нет в словаре, то эта биграмма добавляется.

Следует отметить, что при использовании подобных методов генерации биграмм в словарь могут быть добавлены некоторые биграммы, недопустимые в языке. Однако в силу статистического характера использования словарей биграмм, важно обеспечить лишь то, чтобы шум от несуществующих биграмм, не сводил на нет эффект, получаемый от правильно добавленных биграмм. Использование указанной методики позволило повысить результаты распознавания речи на 1.5-2.0% при исходной надежности распознавания около 95%.

На примере задачи построения информационно-поискового тезауруса и коррекции словаря лексических биграмм лектор старался показать следующие особенности процесса формирования ЛБЗ.

- Процесс формирования ЛБЗ должен быть автоматизированным: он не может быть как полностью ручным (из-за больших объемов ЛБЗ), так и полностью автоматическим (в силу необходимости учета знаний носителей языка и экспертов-лингвистов).
- Для обеспечения высокого качества работы ЕЯ-систем необходимо корректировать статистическую информацию (выявленную при компьютерной обработке текста, например, с помощью методов машинного обучения, и, безусловно, отражающую языковые реалии) с учетом общих лингвистических соображений и явлений, возможно, не проявившихся в исходном корпусе текстов.

Очевидно, что на деятельность разработчика конкретной ЛБЗ могут оказывать существенное влияние временные или экономические (стоимость привлечения высококвалифицированных экспертов) факторы. Однако нельзя забывать, что для создания качественного продукта нужны и время, и знания, и другие ресурсы.

Лингвистический процессор АДАМАНТ.

Общая характеристика ЛП.

Лингвистический процессор (ЛП) АДАМАНТ – адаптивный многоцелевой анализатор текста – разрабатывался в 1986-1990 гг. под руководством М.Г.Мальковского в рамках госбюджетной темы НИР факультета ВМК МГУ "Создание процессора русского языка на базе системы TULIPS-2" (Научно-техническая программа 080.18, Задание 06.01, п. 5.5.4 Координационного плана) и представлял собой один из четырех базовых ЛП для Машинного фонда русского языка.[3]

Процессор был предназначен для анализа фраз естественного (русского) языка и получения их синтактико-семантического представления. Для каждой возможной интерпретации входной фразы процессор строил семантическое представление, соответствующую ему синтаксическую структуру и список отобранных значений словоформ и словосочетаний анализируемой фразы.

Лингвистический процессор АДАМАНТ был ориентирован на автономное использование при решении научных и практических задач, стоящих перед создателями Машинного фонда русского языка. Констатировав этот факт, следует сделать два важных замечания.

Во-первых, участие нашего коллектива в работах по созданию Машинного фонда русского языка, с одной стороны, предъявляло серьезные требования к объему, составу и структуре лингвистического обеспечения ЛП АДАМАНТ (что стимулировало проработку этих вопросов), а, с другой стороны, обеспечивало возможность использования готовых лингвистических описаний, собранных в Машинном фонде.

Первая, но весьма удачная, попытка такого рода взаимодействия была реализована в начале 1990 года, когда наш коллектив получил текст словаря А.А.Зализняка, подготовленный на машинных носителях в ИРЯ АН СССР (головной организации по созданию Машинного фонда русского языка).

Во-вторых, сама идея создания ЛП без проблемно-ориентированных компонентов, казалось бы, вступала в противоречие с основополагающими взглядами авторского коллектива (последовательно проводимыми во всех исследованиях и разработках) на процесс понимания текста, зависимости смысла сообщения от деятельного контекста, от проблемной области. Однако на самом деле такого отрыва не было, поскольку в состав ЛП входили инструментальные блоки МОНИТОР и СЕРВИС, обеспечивавшие настраиваемость ЛП АДАМАНТ на различные области применения в соответствии с предложенными технологическими схемами Эта возможность, кстати, нашла отражение и в самом названии ЛП: адаптивный и многоцелевой.

---

[3] Грандиозный общесоюзный проект по созданию Машинного фонда русского языка предполагал разработку и внедрение разнообразных лингвистических банков данных и ЛП, предназначенных как для автоматизации лингвистических исследований, так и для решения прикладных задач автоматической обработки текста и речи. Первый этап (1986-1990 гг.) был успешно завершен; были определены цели и задачи следующих пятилетних этапов. Однако в 1991 году финансирование проекта, к огромному сожалению, прекратилось по внешним (понятным нам) причинам.

ЛП АДАМАНТ состоял из следующих основных компонентов: лингвистическая база знаний (ЛИНГЗ); модули, непосредственно реализующие анализ входного текста: морфологический, синтаксический и семантический анализаторы (МОРАН, СИНАН, СЕМАН); подсистема языковой адаптациии и обучения (АДАПС), обеспечивавшая возможность обработки новых для ЛП слов и конструкций языка; инструментальные блоки - МОНИТОР и СЕРВИС.

Лингвистические знания. процессора АДАМАНТ (Р-модель) включали описания языковых единиц и правил на лексическом, морфологическом, синтаксическом и семантическом уровнях. В качестве отдельного компонента были выделены лингвистические метазнания, описывающие способы автоматической языковой адаптации и обучения.

Для описания языковых единиц лексического и морфологического уровней была использована полная модель русского словоизменения, базирующаяся на "Грамматическом словаре русского языка" А.А.Зализняка. Эта модель была ориентирована на анализ и синтез словоформ над открытым словарем и обеспечивала возможность автоматического пополнения словаря. Лексические и морфологические компоненты лингвистических знаний процессора содержали: словарь диагностических аффиксов для обработки незнакомых слов, словарь флексий, базовый словарь основ и неизменяемых слов, таблицы чередований и исключений.

Синтаксические и семантические компоненты лингвистических знаний описывали подмножество русского языка, используемое в нескольких узких проблемных областях. На синтаксическом уровне лингвистические знания включали в себя описания: моделей управления синтаксических предикатов, правил грамматики, синтаксических шаблонов, вспомогательный словарь синтаксических групп, словарь фразеологизмов.

Семантические знания процессора были представлены в лексико-семантических зонах словарных статей в виде указателей семантических классов и описаний некоторых сочетаемостных характеристик лексем, а также в словарях семантических предикатов.

Для обмена информацией между блоками анализа процессор использовал предсказания о роли во фразе, строении и значении рассматриваемого фрагмента фразы. Предсказания представляли собой конструкции внутреннего языка процессора, описывающие ожидаемые результаты работы процедур анализа, обеспечивающие целенаправленную обработку входных данных. Аппарат предсказаний играл существенную роль при автоматической языковой адаптации. Предсказания "подсказывали" блоку адаптации каким образом и какие лингвистические знания следует изменить, чтобы завершить анализ непонятного фрагмента фразы.

Основные модули ЛП АДАМАНТ:

Модуль ***МОРАН*** (морфологический уровень). Основные процедуры морфологического анализатора: WANL, MORF, GRAD, STFL.

Модуль ***СИНАН*** (синтаксический уровень). Основные программы, этого уровня: SANL, SAN1, SAN2 и GRR.

Модуль ***СЕМАН***. Формирование семантического представления анализируемой фразы происходило одновременно с анализом ее синтаксической структуры. Ведущей программой синтактико-семантического анализа являлась программа ANAL.

Модуль ***АДАПС***. В конфликтных ситуациях, то есть в случаях, когда имеющаяся совокупность описаний языковых единиц и правил не позволяла установить структуру и/или роль в анализируемой фразе некоторого фрагмента текста лингвистический процессор АДАМАНТ обращался к подсистеме языковой адаптации и обучения АДАПС.

Блок ***МОНИТОР***. Основные функции блока МОНИТОР: организация взаимодействий между программами, участвующими в процессе анализа фразы, и настройка на конкретную (знакомую ЛП) область применения и на конкретный сеанс работы. Такая настройка осуществлялась в режиме администратора или в диалоге с пользователем. В последнем случае МОНИТОР обеспечивал возможность выбора схемы взаимодействия пользователя с ЛП АДАМАНТ. МОНИТОР выяснял, согласен ли пользователь на активное "сотрудничество" с процессором при разрешении конфликтов, возникающих в процессе анализа. В случае его согласия, "самостоятельность" процессора в принятии решений ограничивалась.

Блок ***СЕРВИС***.. В состав блока СЕРВИС входили программы формирования словарей по лингвистическим описаниям, комплекс программ автоматизированного формирования словаря по тексту, а также программы сопровождения словарей, позволяющие группировать словарные статьи слов по различным признакам, упорядочивать имена словарных статей по алфавиту, распечатывать словарь в наглядном виде, синтезировать отдельные формы указанных слов или все их формы.

В заключение остановимся на некоторых технических характеристиках ЛП АДАМАНТ. Базовая его версия была реализована на языке Плэнер для ЭВМ БЭСМ-6. Эксперименты с ЛП АДАМАНТ показали, что возможности системы ПЛЭНЕР-БЭСМ достаточны. Так, при словаре основ объемом 1.5 – 2 тыс. словарных статей использовалось 1-1.2 Мб рабочей памяти, время полной обработки фразы от 1 до 15 сек. (оно зависило от количества незнакомых ЛП слов, в том числе слов с орфографическими ошибками, сложности синтаксической структуры фраз). При работе на ЭВМ "Эльбрус-1КБ" быстродействие увеличивалось в 5 - 7 раз.

Технические характеристики были заметно выше при работе на МВК "Эльбрус-2" в системе программирования ПЛЭНЕР-ЭЛЬБРУС, входной язык которой совместим с входным языком системы ПЛЭНЕР-БЭСМ. Существенно больший объем рабочей памяти позволял работать с более крупными информационными массивами. Скорость работы на МВК "Эльбрус-2" увеличивалась на порядок, а после компиляции основных функций – еще в 3 - 5 раз.

В 1989 году были начаты работы по созданию мобильной версии транслятора с языка Плэнер (версии Плэнер-Паскаль), что позволило вскоре перенести ЛП АДАМАНТ в операционную среду персональных ЭВМ.

Версия ЛП для персональных ЭВМ успешно использовалась для прототипирования нескольких крупных исследовательских и прикладных проектов.

Эксперименты с лингвистическим процессором АДАМАНТ.

*** * * РАБОТАЕТ ЛП "АДАМАНТ" * * ***

КТО?

БОГАЧЕВ

РЕЖИМ РАБОТЫ?

ТЕСТОВЫЙ

ВЫ РАБОТАЕТЕ В ТЕСТОВОМ РЕЖИМЕ!

ЛЮБОЕ ВХОДНОЕ СООБЩЕНИЕ РАССМАТРИВАЕТСЯ ЛИБО КАК КОНТРОЛЬНАЯ ИЛИ ОБУЧАЮЩАЯ ФРАЗА
(НА ЭКРАНЕ РАСПЕЧАТЫВАЕТСЯ НАЙДЕННАЯ ИНТЕРПРЕТАЦИЯ),

ЛИБО КАК ОЦЕНКА ИЛИ УКАЗАНИЕ.

СТЕПЕНЬ САМОСТОЯТЕЛЬНОСТИ СТАНДАРТНАЯ?

ДА

СТЕПЕНЬ ПОДРОБНОСТИ СТАНДАРТНАЯ?

НЕТ

ЗАДАЙТЕ, ПОЖАЛУЙСТА, СТЕПЕНЬ ПОДРОБНОСТИ: ЧИСЛО ОТ 0 - МИН., ДО 5 - МАКС.

4

МОГУ ЛИ Я РАССЧИТЫВАТЬ НА ВАШУ ПОМОЩЬ НА ЭТАПЕ АНАЛИЗА?

ДА

По фамилии пользователя и выбранному им режиму работы ЛП настраивается на очередной сеанс, считывая в свою оперативную память необходимые процедуры и описания уместных в данном сеансе языковых единиц (в частности, словарь пользователя по фамилии Богачев).

Адресуемые данному ("рядовому") пользователю вопросы будут формулироваться в таком виде, что для ответа на них не потребуются знания, выходящие за рамки школьного курса русского языка (при диалоге с администратором в вопросах могут появиться лингвистические термины, понятия языковой модели системы).

В начале сеанса устанавливаются "степень самостоятельности" ЛП при обработке незнакомых языковых объектов и "степень подробности" выдаваемых на экран терминала сообщений о работе процедур языковой адаптации и обучения (этот параметр можно менять в ходе сеанса); ЛП также спрашивает пользователя,

согласен ли он отвечать на его вопросы, возникающие в случаях, когда установленная степень самостоятельности не позволяет процессору изменять лингвистические знания автоматически. При заданной степени подробности (4) процессор будет информировать пользователя о всех обращениях к процедурам языковой адаптации.

ЖДУ ОЧЕРЕДНОЕ СООБЩЕНИЕ

МАЛЬЧИК ИГРАЛ С КОШКОЙ НА БАЛКОНЕ.

ОБРАБАТЫВАЕТСЯ НЕЗНАКОМОЕ СЛОВА "БАЛКОНЕ"

ПРЕДСКАЗАНИЯ: НЕОДУШЕВЛ. СУЩ. В ФОРМЕ ПРЕДЛОЖН. ПАДЕЖА. СЕМ. КЛАСС: =МЕСТО=

УКАЖИТЕ, ПОЖАЛУЙСТА, РОД СУЩЕСТВИТЕЛЬНОГО С ОСНОВОЙ "БАЛКОН"

(ВОЗМОЖНЫЕ ОТВЕТЫ: М/Ж/С)

М

КАКОЕ ОКОНЧАНИЕ ИМЕЕТ СЛОВО "БАЛКОН" В ФОРМЕ ТВОРИТ. ПАДЕЖА ЕД. ЧИСЛА (ОМ ИЛИ ЕМ)?

ОМ

ВСТРЕТИЛОСЬ НЕЗНАКОМОЕ СЛОВО "БАЛКОНЕ"

ОНО ВОСПРИНЯТО КАК НЕОДУШЕВЛ. СУЩ. МУЖ. РОДА В ФОРМЕ ПРЕДЛ. ПАДЕЖА ЕД. ЧИСЛА

ОСНОВЕ ПРИПИСАНЫ ГРАММАТИЧЕСКИЕ ПРИЗНАКИ: НЕОДУШ. СУЩ. МУЖ. РОДА, 1 ТИП СКЛОНЕНИЯ.

ОСНОВЕ ПРИПИСАНЫ СЕМАНТИЧЕСКИЕ ПРИЗНАКИ: =МЕСТО=

ПОНЯТНО (ВРЕМЯ АНАЛИЗА = 3.68750 СЕК.)

=ИГРАТЬ2= (ИГРАТЬ)

=СУБЪЕКТ= : МАЛЬЧИК

=ОБЪЕКТ= : КОШКА

=ВРЕМЯ= : =РАНЕЕ НАСТОЯЩЕГО МОМЕНТА=

=ВИД= : =ДЕЙСТВИЕ ВЫПОЛНЯЛОСЬ=

=МЕСТО= : БАЛКОН (НА)

ПРАВИЛЬНО

НОВАЯ ИНФОРМАЦИЯ ЗАПОМИНАЕТСЯ В ДАННОМ СЕАНСЕ

Предварительный анализ сообщения позволяет предположить, что на вход подано простое предложение.

Анализ входного предложения идет слева направо. Слова поочередно обрабатываются программой/модулем ***МОРАН.***

Слово (точнее, словоформа) *мальчик* системе знакомо; оно состоит из представленной в словаре основы *мальчик*- и пустой флексии -*Ø*. По словарным признакам определяется, что встретилась форма знакомого слова, слово это имеет признаки: одушевленное существительное мужского рода, в форме именительного падежа единственного числа; семантический класс – =ЧЕЛОВЕК=.

Второе слово **–** *играл* – воспринимается как форма известного глагола-предиката *играть* с такими морфо-синтаксическими признаками: единственное число, мужской род, прошедшее время.

В описании глагола *играть* указаны три варианта значения: *играть1* – "играть с кем-то в какую-либо игру"; *играть2* – "играть с кем-то/чем-то"; *играть3* – фразеологизм *играть с огнем* (в значении *рисковать*). Каждому варианту соответствует своя модель управления глагола-предиката *играть*.

Играть2: двухвалентный предикат с актантами: подлежащее и косвенное дополнение. Актант подлежащее в повествовательном предложении обычно (но не обязательно) расположен слева от глагольной формы; он может быть выражен именной группой, главное слово которой (вершина синтаксического дерева) согласуется с глагольной формой по признакам число и род (*мальчик, рыжий человек, мужчина, приехавший к нам вчера из Тамбова, он* и др.*)* , и обладает признаком именительный падеж. Второй актант обычно выражен предложно-именной группой вида: *с/со* + именная группа, главное слово которой (вершина синтаксического дерева) стоит в форме творительного падежа (например, *мальчик играл с той собачкой, которую ему подарили; обезьяна играла с кокосовым орехом*).

На семантическом уровне предикату *играть2* соответствует фрейм-понятие =ИГРАТЬ2= с двумя слотами =СУБЪЕКТ= и =ОБЪЕКТ=.

Играть1: трехвалентный предикат с актантами: подлежащее и два косвенных дополнения. Актант подлежащее, как и в варианте *играть1*, обычно (но не обязательно) расположен слева от глагольной формы и обладает теми же признаками, что и в первом случае. Второй актант выражен предложно-именной группой вида: *с/со* + именная группа, главное слово которой (вершина синтаксического дерева) стоит в форме творительного падежа, например, *со своими постоянными* ***партнерами/с братом*** (вершина дерева выделена полужирным шрифтом). Третий актант выражен предложно-именной группой вида: *в* + именная группа, главное слово которой (вершина синтаксического дерева) стоит в форме винительного падежа, а семантический класс – название некоторой игры (*в* ***преферанс/в настольный теннис***). Второй и третий актанты обычно (при нейтральном порядке слов в повествовательном предложении) обычно находятся в предложении справа от глагольной формы, в произвольном порядке.

На семантическом уровне *играть1* соответствует фрейм-понятие =ИГРАТЬ1= с тремя слотами =СУБЪЕКТ= , =ПАРТНЕР= и =ИГРА=.

Играть3: одновалентный предикат-идиома с актантом подлежащее и "константной" цепочкой слов *с огнем*, расположенной обычно справа от глагольной формы (опять же, при нейтральном порядке слов в повествовательном предложении). Актант подлежащее, как и в варианте *играть1*, обычно (но не обязательно) расположен слева от глагольной формы и обладает теми же признаками, что и в первых двух случаях.

На семантическом уровне *играть3* соответствует фрейм-понятие =РИСКОВАТЬ= с одним слотом =СУБЪЕКТ=.

Обнаружив глагол-предикат анализатор ЛП инициирует процесс ***локально управляемого анализа***, пытаясь для каждого варианта (*играть1, играть2, играть3*) найти как в уже просмотренном фрагменте предложения (*Мальчик*), так и справа от глагола предсказываемые им синтаксические группы или конкретные цепочки слов (*играть* ***с огнем***).

Слова и группы слов, не попавшие в дерево синтаксического анализа и фрейм, описывающий семантику входного сообщения, обрабатываются как потенциальные обстоятельства. Эти действия выполняет специальные программы модулей ***СИНАН*** и ***СЕМАН***.

Если анализируемое предложение удовлетворяет нескольким вариантам, обрабатываются все эти варианты. Однако в качестве первого выдается наиболее вероятный вариант.

Другие варианты могут быть выданы, если пользователь не согласен с первым вариантом анализа (или с последующими). Для этого он может воспользоваться директивой **НЕВЕРНО**, за которой могут следовать те или иные указания (см. пример **МАЛЬЧИК ИГРАЛ С ДРУГОМ В СРЕДУ**).

Обычно предпочтение отдается таким вариантам, которые обеспечивают заполнение наибольшего количества актантов. В данном случае таким вариантом является вариант с двухместным предикатом *играть2*.

Пользователь помогает ЛП определить грамматические признаки незнакомого процессору слова *балкон* (в эксперименте это слово было специально временно удалено из словаря); род и тип склонения (ср. *на кузне*, *на окне*, *на ясене*).

Семантический класс (=МЕСТО=) установлен автоматически, поскольку группа *на балконе* анализировалась процессором после заполнения валентностей выбранной трактовки предиката *играть* (*играть2* – двухместный предикат с актантами: "кто (человек)?", "с кем/чем (животное или предмет)?") – как обстоятельство места (наиболее вероятный тип обстоятельства). Так как пользователь согласился с найденной интерпретацией, ЛП запоминает новую словарную статью слова *балкон*.

Выдаваемый в качестве результата анализа фрейм:

=ИГРАТЬ2= (ИГРАТЬ)

=СУБЪЕКТ= : МАЛЬЧИК

=ОБЪЕКТ= : КОШКА

=ВРЕМЯ= : =РАНЕЕ НАСТОЯЩЕГО МОМЕНТА=

=ВИД= : =ДЕЙСТВИЕ ВЫПОЛНЯЛОСЬ=

=МЕСТО= : БАЛКОН (НА)

содержит:

имя фрейма-понятия – **=ИГРАТЬ2=**;

указанный в скобках фрагмент, выражающий это понятие, – слово **ИГРАТЬ**;

слоты фрейма (**=СУБЪЕКТ=** и **=ОБЪЕКТ=**) с заполняющими соответствующие валентности понятиями: **МАЛЬЧИК и КОШКА**;

далее следуют слоты второго порядка, соответствующие характеристикам, которые неспецифичны для предиката-понятия: значение слота **=ВРЕМЯ=** определяется по прошедшему времени глагола; значение слота **=ВИД=** – по виду (несовершенный вид) и времени глагола; **=МЕСТО=** – по явно указанному в тексте обстоятельству места, выраженному словами *на балконе* (на уровне семантическом – конструкция **БАЛКОН (НА)**).

ЖДУ ОЧЕРЕДНОЕ СООБЩЕНИЕ

МАЛЬЧИК ИГРАЛ С ДРУГОМ В СРЕДУ.

ВСТРЕТИЛОСЬ НЕЗНАКОМОЕ СЛОВО "СРЕДУ"

ОНО ВОСПРИНЯТО КАК НЕОДУШ. СУЩ. ЖЕНСКОГО РОДА В ФОРМЕ ВИНИТ. ПАДЕЖА ЕД. ЧИСЛА

ОСНОВЕ ПРИПИСАНЫ ГРАММАТИЧЕСКИЕ ПРИЗНАКИ: НЕОДУШ. СУЩ. ЖЕНСКОГО РОДА, 2 ТИП СКЛОНЕНИЯ

ОСНОВЕ ПРИПИСАНЫ СЕМАНТИЧЕСКИЕ ПРИЗНАКИ: =ИГРА=

ПОНЯТНО (ВРЕМЯ АНАЛИЗА = 3.10875 СЕК.)

=ИГРАТЬ1= (ИГРАТЬ)

=СУБЪЕКТ= : МАЛЬЧИК

=ПАРТНЕР= : ДРУГ

=ИГРА= : СРЕДА

=ВРЕМЯ= : =РАНЕЕ НАСТОЯЩЕГО МОМЕНТА=

=ВИД= : =ДЕЙСТВИЕ ВЫПОЛНЯЛОСЬ=

НЕВЕРНО. СРЕДА - ДЕНЬ НЕДЕЛИ

УКАЗАНИЕ УЧТЕНО. НОВАЯ ИНТЕРПРЕТАЦИЯ:

=ИГРАТЬ1= (ИГРАТЬ)

=СУБЪЕКТ= : МАЛЬЧИК

=ПАРТНЕР= : ДРУГ

=ИГРА= : ?

=ВРЕМЯ= : =РАНЕЕ НАСТОЯЩЕГО МОМЕНТА= : СРЕДА

=ВИД= : =ДЕЙСТВИЕ ВЫПОЛНЯЛОСЬ=

ВЕРНО. РЕЖИМ: СП = 2.

НОВАЯ ИНФОРМАЦИЯ ЗАПОМИНАЕТСЯ В ДАННОМ СЕАНСЕ

ИЗМЕНЕН ПАРАМЕТР: СТЕПЕНЬ ПОДРОБНОСТИ (СП)

Попытка процессора воспринять группу *в среду* как один из актантов трехместного предиката *играть1* ("кто (человек)?", "с кем (человек)?", "во что (игра)?") оказалась неудачной. Стратегия предпочтения варианта, обеспечивающего заполнение максимального количества актантов, в данном случае "подвела" ЛП. Причина в том, что, опять же, в целях эксперимента из словаря было временно удалено слово *среда*.

Заметим, однако, что в русском языке возможны корректные варианты интерпретаций близких по словарному составу предложений, например, *играл в чехарду*.

Пользователь отверг найденную интерпретацию и определил семантику незнакомого слова *среда* через знакомое ЛП понятие *день недели*. верного результата можно был бы добиться и другими способами.

После этого процессор: проанализировал словоформу *среда* с предсказанием "существительное в форме именительного падежа, семантический класс =ВРЕМ1="; отменил все действия, выполненные с момента начала обработки группы *в среду* (группы *мальчик* и *с другом*, выбранные на роли первого и второго актантов, повторно не обрабатывались); продолжил анализ фразы. Слово *среда* теперь – по семантическим признакам – не может быть воспринято как название игры. Третья валентность предиката *играть1* остается незаполненной, а на этапе поиска обстоятельств группа *в среду* воспринимается как обстоятельство времени.

Согласившись с новой интерпретацией пользователь (с помощью директивы **РЕЖИМ**) уменьшает значение параметра "степень подробности". Поэтому в дальнейшем о незнакомых словах выдается менее подробная информация.

ЖДУ ОЧЕРЕДНОЕ СООБЩЕНИЕ

МАЛЬЧИК СТРИГЕТ КУЗЯВУЮ БУТЯВКУ НА БАЛКНЕ.

"СТРИГЕТ" - ОШИБКА В СПРЯЖЕНИИ! ПРИ ФЛЕКСИИ "ЕТ" В ОСНОВЕ ПРОИСХОДИТ ЧЕРЕДОВАНИЕ: Г - Ж. ОШИБКА ИСПРАВЛЕНА

КАКАЯ ИЗ ОСНОВ: 1. БАЛКОН, 2. БАЛКН - ВВЕДЕНА ВЕРНО (1 ИЛИ 2, 0 - ЕСЛИ ОБЕ ВВЕДЕНЫ ВЕРНО)?

1

ВСТРЕТИЛОСЬ НЕЗНАКОМОЕ СЛОВО "БУТЯВКУ"

- ОДУШ. СУЩ. ЖЕН. РОДА В ФОРМЕ ВИНИТ. ПАДЕЖА ЕД. ЧИСЛА (13 ТИП СКЛОНЕНИЯ)

ВСТРЕТИЛОСЬ НЕЗНАКОМОЕ СЛОВО "КУЗЯВУЮ"

- ПРИЛ. В ФОРМЕ ВИН. ПАДЕЖА ЕД. ЧИСЛА (ЖЕНСКИЙ РОД) (1 ТИП СКЛОНЕНИЯ)

ПОНЯТНО (ВРЕМЯ АНАЛИЗА = 4.80000 СЕК.)

=СТРИЧЬ= (СТРИЧЬ)

=СУБЪЕКТ= : МАЛЬЧИК

=ОБЪЕКТ= : БУТЯВКА --АТР -> КУЗЯВАЯ

=ВРЕМЯ= : =СЕЙЧАС=/=ВСЕГДА=/=ИНОГДА=

=ВИД= : =ДЕЙСТВИЕ ВЫПОЛНЯЕТСЯ=/=ДЕЙСТВИЕ ХАРАКТЕРНО=

ВЕРНО

НОВАЯ ИНФОРМАЦИЯ ЗАПОМИНАЕТСЯ В ДАННОМ СЕАНСЕ

При установленных значениях параметров процессор исправляет грамматические и случайные ошибки в знакомых словах самостоятельно (информируя об этом пользователя). Исправление случайной ошибки в слове *балкон*, впервые появившемся только в данном сеансе, возможно только с согласия пользователя. Поэтому АДАМАНТ и обращается к пользователю:

КАКАЯ ИЗ ОСНОВ: 1. БАЛКОН, 2. БАЛКН - ВВЕДЕНА ВЕРНО (1 ИЛИ 2, 0 - ЕСЛИ ОБЕ ВВЕДЕНЫ ВЕРНО)?.

В данном предложении встретилась именная группа *кузявую бутявку* (вида <прилагательное> + <существительное>, согласующиеся по значениям признаков: род, число, падеж). Хотя оба этих слова системе незнакомы, они получают естественную интерпретацию: *кузявую* – **ПРИЛ. В ФОРМЕ ВИН. ПАДЕЖА ЕД. ЧИСЛА (ЖЕНСКИЙ РОД) (1 ТИП СКЛОНЕНИЯ)**, *бутявку* – **- ОДУШ. СУЩ. ЖЕН. РОДА В ФОРМЕ ВИНИТ. ПАДЕЖА ЕД. ЧИСЛА (13 ТИП СКЛОНЕНИЯ)**, при этом слово *кузявая* трактуется как атрибут/признак слова *бутявка*.

К запоминанию в словаре подготовлены: основы прилагательного *кузяв-* и существительного *бутявк-* (с соответствующими грамматическими признаками). После того, как пользователь одобрил результаты анализа (директива **ВЕРНО**), окончательно формируются и вносятся в словарь соответствующие словарные статьи.

Отметим, что правильной интерпретации этой именной группы помогло то, что поиск ее (локально-управляемый анализ) опирался на предсказания знакомого слова-предиката *стричь* (напомним, что ошибку в чередовании *стригет* вместо *стрижет* – процессор исправил самостоятельно). Глагол *стричь* описан в словаре как трехместный предикат *стричь* (с актантами: "кто (человек)?", "кого (человек/животное)?", "инструмент").

ЖДУ ОЧЕРЕДНОЕ СООБЩЕНИЕ

ГЛОКАЯ КУЗДРА ШТЕКО БУДЛАНУЛА БОКРА И КУДРЯЧИТ БОКРЕНКА.

ОБРАБАТЫВАЕТСЯ НЕЗНАКОМОЕ СЛОВО "БУДЛАНУЛА"

ЭТО НЕЗНАКОМОЕ СЛОВО ТРАКТУЕТСЯ КАК ГЛАГОЛ-ПРЕДИКАТ. ВЫ СОГЛАСНЫ (ДА, НЕТ)?

ДА

ОБРАБАТЫВАЕТСЯ НЕЗНАКОМОЕ СЛОВО "КУДРЯЧИТ"

ЭТО НЕЗНАКОМОЕ СЛОВО ТРАКТУЕТСЯ КАК ГЛАГОЛ-ПРЕДИКАТ. ВЫ СОГЛАСНЫ (ДА, НЕТ)?

ДА

При анализе этой знаменитой фразы процессор по синтаксическим шаблонам выделяет в ее составе два ядерных предложения, связанных союзом *и*. В каждом из них на роли слов-предикатов выбираются (с согласия пользователя) формы незнакомых глаголов: *будланyть* и *кудрячить*. Этим глаголам сопоставляется одна из простейших стандартных моделей управления ("кто?", "кого/что?"), которая используется при поиске актантов. На роль подлежащего второго предложения выбирается именная группа *глокая куздра* – подлежащее первого предложения. После уточнения грамматических признаков нескольких незнакомых слов распечатывается полученная интерпретация.

СУЩЕСТВИТЕЛЬНОЕ "БОКР" - ОДУШЕВЛЕННОЕ (ДА, НЕТ)?

ДА

КАК ПРАВИЛЬНО: 1. ГЛОКИЙ ИЛИ 2. ГЛОКОЙ?

1

СУЩЕСТВИТЕЛЬНОЕ "БОКРЕНОК" - ОДУШЕВЛЕННОЕ (ДА, НЕТ)?

ДА

ВСТРЕТИЛОСЬ НЕЗНАКОМОЕ СЛОВО "ШТЕКО" - НАРЕЧИЕ (ОБР. ДЕЙСТВИЯ)

ВСТРЕТИЛОСЬ НЕЗНАКОМОЕ СЛОВО "БОКРА"
- ОДУШ. СУЩ. МУЖ. РОДА В ФОРМЕ ВИНИТ. ПАДЕЖА ЕД. ЧИСЛА (1 ТИП СКЛОНЕНИЯ)

ВСТРЕТИЛОСЬ НЕЗНАКОМОЕ СЛОВО "КУЗДРА"
- ОДУШ. СУЩ. ЖЕН. РОДА В ФОРМЕ ИМ. ПАДЕЖА ЕД. ЧИСЛА (2 ТИП СКЛОНЕНИЯ)

ВСТРЕТИЛОСЬ НЕЗНАКОМОЕ СЛОВО "ГЛОКАЯ"
- ПРИЛ. В ФОРМЕ ИМ. ПАДЕЖА ЕД. ЧИСЛА (ЖЕНСКИЙ РОД) (3 ТИП СКЛОНЕНИЯ)

ВСТРЕТИЛОСЬ НЕЗНАКОМОЕ СЛОВО "БУДЛАНУЛА"
- ГЛАГОЛ (НЕВОЗВР., СОВ. ВИД) В ФОРМЕ: ПРОШ. ВР., ЖЕН. РОД, ЕД. ЧИСЛО (2 ТИП СПР.)

ВСТРЕТИЛОСЬ НЕЗНАКОМОЕ СЛОВО "БОКРЕНКА"

- ОДУШ. СУЩ. МУЖ. РОДА В ФОРМЕ ВИН. ПАДЕЖА ЕД. ЧИСЛА (3 ТИП СКЛОНЕНИЯ);
ОСНОВА МНОЖ. ЧИСЛА: "БОКРЯТ"

ВСТРЕТИЛОСЬ НЕЗНАКОМОЕ СЛОВО "КУДРЯЧИТ"
- ГЛАГОЛ (НЕВОЗВР., НЕСОВ. ВИД) В ФОРМЕ: НАСТ. ВР., 3 ЛИЦО, ЕД. ЧИСЛО (4 ТИП СПР.)

ПОНЯТНО (ВРЕМЯ АНАЛИЗА = 13.78025 СЕК.)

=ВЫПОЛНЯТЬ ДЕЙСТВИЕ= (БУДЛАНУТЬ)
=СУБЪЕКТ= : КУЗДРА --АТР -> ГЛОКАЯ
=ОБЪЕКТ= : БОКР
=ВРЕМЯ= : =РАНЕЕ НАСТОЯЩЕГО МОМЕНТА=
=ВИД= : =ДЕЙСТВИЕ ВЫПОЛНЕНО=
=ОБР. Д.= : ШТЕКО
======ЗАТЕМ==> =ВЫПОЛНЯТЬ ДЕЙСТВИЕ= КУДРЯЧИТЬ)
=СУБЪЕКТ= : КУЗДРА --АТР -> ГЛОКАЯ
=ОБЪЕКТ= : БОКРЕНОК
=ВРЕМЯ= : =СЕЙЧАС=
=ВИД= : =ДЕЙСТВИЕ ВЫПОЛНЯЕТСЯ=

ВЕРНО

НОВАЯ ИНФОРМАЦИЯ ЗАПОМИНАЕТСЯ В ДАННОМ СЕАНСЕ

ЖДУ ОЧЕРЕДНОЕ СООБЩЕНИЕ

ВЧЕРА КУЗДРА БУДЛАНУЛА БОКРА И ШТЕКО БАРАБАРДАЕТСЯ.

ОБРАБАТЫВАЕТСЯ НЕЗНАКОМОЕ СЛОВО "БАРАБАРДАЕТСЯ"

ЭТО НЕЗНАКОМОЕ СЛОВО ТРАКТУЕТСЯ КАК ГЛАГОЛ-ПРЕДИКАТ. ВЫ СОГЛАСНЫ (ДА, НЕТ)?

ДА

ВСТРЕТИЛОСЬ НЕЗНАКОМОЕ СЛОВО "БАРАБАРДАЕТСЯ"
ОНО ВОСПРИНЯТО КАК ГЛАГОЛ (ВОЗВР., НЕСОВ. ВИД): НАСТ. ВР., 3 ЛИЦО, ЕД. ЧИСЛО (1 ТИП СПР.)

ПОНЯТНО (ВРЕМЯ АНАЛИЗА = 4.53125 СЕК.)

=ВЫПОЛНЯТЬ ДЕЙСТВИЕ= (БУДЛАНУТЬ)
=СУБЪЕКТ= : КУЗДРА
=ОБЪЕКТ= : БОКР

=ВРЕМЯ= : ИЮНЬ --ДЕНЬ -> 5

=ВИД= : =ДЕЙСТВИЕ ВЫПОЛНЕНО=

======ЗАТЕМ==>

=ВЫПОЛНЯТЬ ДЕЙСТВИЕ= (БАРАБАРДАТЬСЯ)

=СУБЪЕКТ= : КУЗДРА

=ОБЪЕКТ= : КУЗДРА

=ВРЕМЯ= : =СЕЙЧАС=

=ВИД= : =ДЕЙСТВИЕ ВЫПОЛНЯЕТСЯ=

=ОБР. Д.= : ШТЕКО

ВЧЕРА КУЗДРА БУДЛАНУЛА БОКРА, И БУТЯВКА БАРАБАРДАЕТСЯ.

ПОСКОЛЬКУ ОЦЕНКИ НЕ БЫЛО, СЧИТАЕТСЯ, ЧТО ПРЕДЫДУЩАЯ ФРАЗА ПОНЯТА ПРАВИЛЬНО

АНАЛИЗ ОЧЕРЕДНОГО СООБЩЕНИЯ:

ПОНЯТНО (ВРЕМЯ АНАЛИЗА = 4.01250 СЕК.)

=ВЫПОЛНЯТЬ ДЕЙСТВИЕ= (БУДЛАНУТЬ)

=СУБЪЕКТ= : КУЗДРА

=ОБЪЕКТ= : БОКР

=ВРЕМЯ= : ИЮНЬ --ДЕНЬ -> 5

=ВИД= : =ДЕЙСТВИЕ ВЫПОЛНЕНО=

========И1 ==>

=ВЫПОЛНЯТЬ ДЕЙСТВИЕ= (БАРАБАРДАТЬСЯ)

=СУБЪЕКТ= : БУТЯВКА

=ОБЪЕКТ= : БУТЯВКА

=ВРЕМЯ= : =СЕЙЧАС=

=ВИД= : =ДЕЙСТВИЕ ВЫПОЛНЯЕТСЯ=

ВЕРНО

ЖДУ ОЧЕРЕДНОЕ СООБЩЕНИЕ

ГЛОКАЯ КУЗДРА, ШТЕКО БУДЛАНУЛА БОКРА.

ФРАЗА НЕПОНЯТНА (ВРЕМЯ АНАЛИЗА = 3.73025 СЕК.)

НЕСВЯЗНЫЕ ФРАГМЕНТЫ:

1) КУЗДРА --АТР -> ГЛОКАЯ

2) =ВЫПОЛНЯТЬ ДЕЙСТВИЕ= (БУДЛАНУТЬ)

=СУБЪЕКТ= : ?

=ОБЪЕКТ= : БОКР

=ВРЕМЯ= : =РАНЕЕ НАСТОЯЩЕГО МОМЕНТА=
=ВИД= : =ДЕЙСТВИЕ ВЫПОЛНЕНО=
=ОБР. Д.= : ШТЕКО

ЗАБЫТЬ: ГЛОКАЯ, КУЗДРА, ШТЕКО, БУДЛАНУТЬ, БОКР

СДЕЛАНО

РЕЖИМ: СП = 0

ИЗМЕНЕН ПАРАМЕТР: СТЕПЕНЬ ПОДРОБНОСТИ (СП)

ЖДУ ОЧЕРЕДНОЕ СООБЩЕНИЕ

ГЛОКАЯ КУЗДРА, ШТЕКО БУДЛАНУЛА БОКРА.

ПОНЯТНО (ВРЕМЯ АНАЛИЗА = 8.01250 СЕК.)

=ВЫПОЛНЯТЬ ДЕЙСТВИЕ= (БУДЛАНУТЬ)
=СУБЪЕКТ= : ШТЕКО
=ОБЪЕКТ= : БОКР
=ВРЕМЯ= : =РАНЕЕ НАСТОЯЩЕГО МОМЕНТА=
=ВИД= : =ДЕЙСТВИЕ ВЫПОЛНЕНО=
======ПРИ ЭТОМ==>
=ВЫПОЛНЯТЬ ДЕЙСТВИЕ= (ГЛОКАТЬ)
=СУБЪЕКТ= : ШТЕКО
=ОБЪЕКТ= : КУЗДР
=ВРЕМЯ= : =РАНЕЕ НАСТОЯЩЕГО МОМЕНТА=
=ВИД= : =ДЕЙСТВИЕ ВЫПОЛНЯЛОСЬ=

ВЕРНО. КОНЕЦ**.**

НОВАЯ ИНФОРМАЦИЯ ЗАПОМИНАЕТСЯ В ДАННОМ СЕАНСЕ

РАБОТАЕТ БЛОК ОБУЧЕНИЯ

СЕАНС ОКОНЧЕН

Последние два примера показывают, что ЛП АДАМАНТ может забывать (директива **ЗАБЫТЬ: <список слов в начальной форме, разделенных запятыми>**) информацию, полученную в ходе сеанса, и что результат анализа фразы зависит от наличных (и меняющихся во времени) лингвистических знаний.

После получения правильной трактовки (одобренной пользователем и согласующейся с общепринятыми) знаменитой фразы академика Льва Владимировича Щербы **ГЛОКАЯ КУЗДРА ШТЕКО БУДЛАНУЛА БОКРА И КУДРЯЧИТ БОКРЕНКА** система запомнила ранее неизвестные ей (не входящие в словарь):

наречие **ШТЕКО** (как неизменяемое слово, тип наречия (образа действия) был установлен из соображений частотности употребления обстоятельственных наречий,

существительное **БОКР** (как начальную форму существительного **БОКРА**) с установленными – в ходе анализа с предсказаниями – грамматическими признаками, характеризующими склонение этого слова,

существительное **КУЗДРА** (как начальную форму существительного **КУЗДРА**) – опять же с описанием склонения этого слова,

глагол **БУДЛАНУТЬ** (как начальную форму глагола **БУДЛАНУЛА**) – глагола-предиката с двумя актантами ("кто?", "кого?") и описанием особенностей спряжения,

глагол **КУДРЯЧИТЬ** (как начальную форму глагола **КУДРЯЧИТ**) – также глагола-предиката с двумя актантами ("кто?", "кого?") и своими особенностями спряжения,

прилагательное **ГЛОКИЙ** (как начальную форму прилагательного **ГЛОКАЯ**),

существительное **БОКРЕНКА** (как начальную форму существительного **БОКРЕНОК**) с установленными – в ходе анализа с предсказаниями – грамматическими признаками, характеризующими склонение этого слова и с учетом строения слова (особенности образования форм множественного числа – от основы **БОКРЯТ**).

Следующие два предложения:
ВЧЕРА КУЗДРА БУДЛАНУЛА БОКРА И ШТЕКО БАРАБАРДАЕТСЯ.
ВЧЕРА КУЗДРА БУДЛАНУЛА БОКРА, И БУТЯВКА БАРАБАРДАЕТСЯ.
– не противоречат построенным описаниям и обрабатываются без всяких заминок (и быстро, поскольку все слова теперь ЛП знакомы).

Однако предложение: **ГЛОКАЯ КУЗДРА, ШТЕКО БУДЛАНУЛА БОКРА.** – при сформированных описаниях всех пяти слов проанализировано быть не может. Продолжение диалога с ЛП показывает один из возможных выходов из этой ситуации. Директива **ЗАБЫТЬ: ГЛОКАЯ, КУЗДРА, ШТЕКО, БУДЛАНУТЬ, БОКР** превращает эти слова в новые (сформированные ранее словарные статьи забываются). В результате строится полная интерпретация данного предложения (разумеется, с новыми признаками входящих в него слов). Никакой информации о появлении и свойствах новых слов не печатается, так как пользователь выбрал (директива **РЕЖИМ: СП = 0**) минимально возможный уровень подробности сообщений системы.

Завершая сеанс, ЛП запоминает новую информацию в своей долговременной памяти. В данном случае (в словарь пользователя Богачева) будут записаны словарные статьи слов: *балкон, куздр*, *глокать* (глагол, имеющий форму деепричастия *глокая*) и др. в соответствии с последним одобренным пользователем результатом анализа. Разумеется, запоминаются одобренные результаты анализа предложений, предшествующих "*глокой куздре*".

Система комплексного контроля качества текста ЛИНАР

Функции системы ЛИНАР; сценарии работы с системой

Построение автокорректоров сталкивается с рядом принципиальных и не решенных пока в полном объеме проблем: компактное хранение словарей, эффективные методы морфологического и синтаксического анализа и т.д. Тем не менее на очереди - создание систем, способных производить более сложное по сравнению с автокорректорами автоматическое или автоматизированное редактирование текстов на естественном языке. В идеале же необходима система, выполняющая функции научного редактора - человека, осуществляющего литературную и научную правку научно-технических текстов. Такое направление развития представляет разрабатывавшаяся в 1986-1990 гг. на кафедре алгоритмических языков факультета ВМК МГУ система ЛИНАР (ЛИтературно-НАучный Редактор) - интеллектуальная система комплексного контроля качества и редактирования русскоязычных текстов. Интересна система и в плане ***Работы на ограниченном языке*** (Лекция 11).

Суть подхода заключалась в существенном расширении возможностей имевшихся в то время автокорректоров за счет:
- ограничения предметной области, к которой относились обрабатываемые тексты (методы, алгоритмы и программы обработки данных телеметрии на многопроцессорных вычислительных комплексах);
- ограничения видов текстов (научно-технические отчеты, деловая переписка);
- использования средств синтаксического и семантического анализа текста;
- привлечения более полных моделей русского языка.

Пользователем ЛИНАР является человек, оценивающий с помощью системы качество некоторого текста с позиций лица, которому адресован этот текст (адресата), и вносящий в текст необходимые исправления. В качестве адресата могут выступать литературный или научный редактор, корректор, потенциальные читатели (конструкторы, программисты, руководители). Пользователем ЛИНАР может быть, например, автор обрабатываемого текста, желающий взглянуть на него «со стороны», или научный руководитель работы, обеспокоенный терминологическими и стилистическими неувязками в текстах разделов, подготовленных различными участниками проекта.

Обработка текста с помощью системы ЛИНАР включает в себя в общем случае несколько циклов (как и при подготовке текста «вручную»), каждый из которых оформляется как самостоятельный сеанс работы с системой. В начале сеанса пользователь формирует задание на обработку текста, для выполнения которого система загружает необходимые информационные модули и вызывает программы контроля текста. Каждая программа проверяет некоторое определенное свойство текста, т.е. реализует одноаспектный контроль текста. Таким образом, в структурном плане систему ЛИНАР можно считать пакетом прикладных программ; сеанс работы с ней состоит из серии одноаспектных проверок текста или его фрагментов.

Основная технологическая схема использования системы ЛИНАР предусматривает, что текст хранится на машинных носителях и обрабатывается программами контроля, формирующими протокол замечаний по тексту (иногда система предлагает свой вариант исправления). Далее пользователь просматривает эти замечания и, если он с ними соглашается, вносит необходимые изменения в текст с помощью текстового редактора. Измененная версия текста может быть объектом обработки в следующем сеансе. В зависимости от объема текста пользователь может выбрать диалоговый или пакетный

режим работы с системой. В последнем случае протокол замечаний формируется на внешнем носителе.

Отметим, что используемые в ЛИНАР знания позволяют системе фиксировать различные типы конфликтных ситуаций (и формировать соответствующие замечания). Однако как бы полны ни были знания ЛИНАР, обнаружить все неточности, противоречия, неопределенности система самостоятельно не может. Поэтому часть программ контроля собирает некоторую вспомогательную информацию о тех или иных характеристиках (свойствах) текста, не давая ей оценки.

Например, при написании отдельных фрагментов текста разными авторами для обозначения одной и той же сущности могут быть использованы различные термины, что усложняет понимание текста. Автоматическое обнаружение подобных конфликтов требует привлечения глубоких знаний о понятийном и терминологическом аппарате предметной области, и в ЛИНАР не реализуется. Однако в составе системы имеется программа контроля, которая может сформировать по фрагментам текста списки используемых терминологических словосочетаний. На основе этой информации решить терминологические проблемы человеку будет значительно проще, чем при обработке текста «вручную».

ЛИНАР не только обнаруживает неточности, ошибки, но и может «объяснить» пользователю суть своих замечаний, а также предложить способы устранения ошибок. Так, например, в случае орфографической ошибки система предлагает свой вариант исправления слова, в случае нарушения естественного порядка слов - правильный порядок слов и т.д. Рекомендации системы призваны помочь пользователю в улучшении текста, направляют его деятельность.

База знаний системы

Контроль текста, осуществляемый системой ЛИНАР, основывается на использовании знаний о том, что такое правильный, хороший текст. Совокупность этих знаний называется контролирующими знаниями, или К-знаниями. При формировании К-знаний учитывались результаты лингвистических, психологических работ, исследований по эргономике; принят во внимание опыт редакторов, корректоров, нормоконтролеров.

К-знания должны обеспечить возможность оценки текста с различных сторон:

- соответствие общеязыковым нормам;

- соответствие «внешним» нормам, например, требованиям ГОСТов, регламентирующих форму изложения материала в научно-технических документах;

- сложность восприятия текста потенциальным читателем;

- семантическая корректность текста (соответствие выявляемых в тексте семантических отношений и понятийной модели предметной области).

Часть К-знаний (процедурная составляющая) представлена программами одноаспектного контроля. Каждая программа фиксирует строго определенное свойство текста или строго определенный дефект текста (конфликтную ситуацию). Затем формируется соответствующее диагностическое сообщение, которое, в зависимости от выбранного режима работы, либо сразу предъявляется пользователю, либо включается в протокол замечаний.

Важным компонентом информационного обеспечения системы ЛИНАР является и лингвистическая база знаний, содержащая базовые общие знания о русском языке. Кроме

того, ЛИНАР использует тематический словарь и тезаурус предметной области, к которой относятся обрабатываемые тексты, и описания нормативных требований, предъявляемых к текстам. Соответствующие информационные массивы создавались разработчиками системы на основе общеязыковых и предметно-ориентированных словарей и справочников, Государственных стандартов и отраслевых инструкций по оформлению текстовых документов.

База знаний ЛИНАР содержит также заранее формируемый - и пополняемый в ходе эксплуатации системы - **банк адресатов**: конкретных читателей или определенных однородных групп читателей (конкретный руководитель научно-исследовательского проекта; конкретный представитель руководства организации-заказчика; инженеры, которые будут создавать описываемый программно-аппаратный комплекс и др.). Настройка на адресата производится в начале очередного сеанса работы с ЛИНАР. При такой настройке могут меняться базовые и тематические лингвистические знания (состав словаря, совокупность грамматических правил), степень жесткости требований по соблюдению тех или иных норм и условий.

Чтобы задать эту информацию, следует указать имя одного из известных ЛИНАР адресатов (или идентификатор известной группы адресатов) и выбрать значения дополнительных параметров программ контроля.

С помощью такой настройки удается моделировать процесс восприятия текста разными адресатами и, следовательно, оценивать качество текста с разных точек зрения.

Таким образом, К-знания ЛИНАР (которые служат критерием корректности текста и используются для обнаружения «дефектов» текста - отклонений от требований, предъявляемых К-знаниями) формируются динамически в каждом конкретном сеансе работы с системой и являются комплексными по своей природе. Они включают как процедурные знания об исследуемом аспекте текста (воплощенные в соответствующих программах контроля), так и декларативные знания, фильтруемые и конкретизируемые в начале каждого сеанса.

Обнаруженные программой контроля несоответствия текста и К-знаний могут быть устранены двумя способами:

- путем внесения изменений в текст (это наиболее частый случай: несоответствие - суть ошибка, допущенная в тексте, которую необходимо исправить);
- путем изменения К-знаний системы.

Заметим, что изменениям подвергается лишь один компонент К-знаний - лингвистические знания, причем не все, а лишь те, которые соответствуют наиболее подвижной части естественного языка - лексикону. Как правило, такие изменения заключаются в пополнении базы знаний, например, в создании новой словарной статьи для слова, впервые встретившегося в тексте и не знакомого системе.

Знания, отображающие требования семантической корректности и простоты интерпретации, общеязыковые и внешние нормы, может изменять только администратор системы.

Для внесения изменений в базу лингвистических знаний используются сервисные программы; для изменения текста - подсистема редактирования ЛИНАРа.

Отметим, что (даже при работе с ЛИНАР в диалоговом режиме) редактирование текста обычно производится по завершении работы программ контроля. Это связано с

тем, что исправление фиксируемых системой ошибок и неточностей зачастую требует переделки относительно больших фрагментов текста (разбиение длинной фразы на несколько более простых, устранение неоднозначности трактовки и т.п.). Однако некоторые - локальные - изменения можно внести в текст сразу же в момент обнаружения ошибки. Поэтому в ряде программ контроля, например, в программах орфографического уровня, предусмотрена возможность исправления фиксируемых ошибок в момент их обнаружения.

Программы контроля

Программы контроля текста могут быть классифицированы по нескольким критериям.

Первый критерий связан с анализируемым программой аспектом текста. В соответствии с этим критерием выделяются следующие группы программ одноаспектного контроля:

- контроль орфографии (включая поиск ошибок в склонении и спряжении слов);
- анализ лексического состава текста;
- стилистический контроль;
- проверка выполнения правил структуризации текста;
- контроль синтаксической структуры;
- пунктуационный контроль;
- семантический контроль.

По второму критерию программы одноаспектного контроля подразделяются на программы локального и глобального анализа текста. Программы первой группы обрабатывают мелкие фрагменты текста: отдельные словоформы, словосочетания, специальные символы, не исследуя их контекстные связи или ограничиваясь учетом ближайшего окружения (соседнего слова справа, например). Локальный анализ характерен для программ орфографического, лексического и (частично) стилистического контроля. Программы, осуществляющие глобальный анализ, исследуют, как правило, структуру более крупных единиц текста: фраз и иногда абзацев (синтаксический и семантический контроль), текста в целом.

Третий критерий связан с характером результата, получаемого программой одноаспектного анализа. Основная часть программ контроля обнаруживает те или иные несоответствия текста и К-знаний, используемых в текущем сеансе. Результатом их работы является список выявленных несоответствий (нарушений). Однако некоторые программы, как уже отмечалось, определяют отдельные свойства текста, не оценивая их. Так, программа ЛЕКС1 составляет частотный словарь исследуемого текста (фрагмента текста). Оценку полученным результатам дает человек - пользователь ЛИНАР, он же принимает решение о дальнейших действиях. Его реакция может быть, например, такой - поработать над текстом пункта 4.5.1., поскольку в этом тексте (занимающем всего две страницы) 26 раз встречается слово *знания* (в различных формах) и 7 раз - слово *соответственно*.

Только что рассмотренный пример (программа ЛЕКС1) можно использовать и для иллюстрации четвертого критерия классификации программ контроля. Эта программа, как и ряд других, выдает некоторую глобальную информацию об исследуемом фрагменте текста, не фиксируя, в каких позициях (абзацах, фразах или строках) были обнаружены в

тексте формы различных слов. Другие программы, например программы проверки орфографии, локализуют обнаруживаемые ими свойства (дефекты) текста.

И наконец, отметим еще одно (формальное) различие программ контроля. Для всех программ основным параметром является подлежащий обработке фрагмент текста. Однако для некоторых программ нужно обязательно указать дополнительные параметры, конкретизирующие задание. Например, при вызове программы ЛЕКС2 нужно указать, какие именно грамматические признаки слов интересуют пользователя.

Некоторые программы контроля получают в качестве параметра предельно допустимые (пороговые) числовые значения количественно оцениваемых параметров текста. Отметим, что, меняя порог, можно варьировать уровень требований, предъявляемых к тексту, моделируя тем самым оценку его разными адресатами. Например, можно установить в качестве предельно допустимой длины фразы 25 слов или ограничить число придаточных предложений (в составе сложного предложения) двумя. Фразы, в которых эти пороговые значения превышены, будут классифицированы соответствующими программами контроля как недопустимые.

Орфографический контроль

Программы орфографического контроля обнаруживают (и предлагают варианты исправления) мотивированные грамматические ошибки в основах и окончаниях (флексиях) слов, записанных в словарь системы, и слов, встретившихся ей впервые (незнакомых), а также случайные, или немотивированные, ошибки.

Основные классы учитываемых случайных ошибок таковы:

- пропуск одной буквы (*а**с**емблер*),

- одна лишняя буква (*авт**т**окод*),

- замена одной буквы (*ко**н**пьютер*),

- перестановка двух соседних букв (*аг**л**оритм*).

Признаком ошибки часто служит появление в обрабатываемом тексте формы незнакомого слова.

Предпринимается попытка «свести» такое незнакомое слово к знакомому с помощью преобразований, обратных перечисленным выше (считается, что ошибка могла возникнуть в результате одного из таких «прямых» преобразований знакомого слова). Для предварительной оценки близости слов (основ слов) используется специально разработанная метрика.

Одна из программ обнаруживает ошибки в датах, задаваемых в тексте с помощью конструкций вида ДД.ММ.ГГ. Если задан и диапазон возможных дат, проверяется также принадлежность всех представленных в исследуемом тексте дат этому диапазону.

Примеры работы программ:

прочитанна - ОШИБКА В СЛОВОИЗМЕНЕНИИ !
ОЖИДАЕМОЕ СЛОВО: прочитана

рассчета - ВОЗМОЖНА ОШИБКА ТИПА "удвоение буквы"
ОЖИДАЕМОЕ СЛОВО : расчета

10.25.89.
ОШИБКА В ДАТЕ - недопустимая дата: месяц: 25

Анализ лексического состава текста

Программа ЛЕКС1

Программа подсчитывает, сколько раз в тексте (области) употребляется то или иное слово. Программа формирует полный список всех различных слов текста с указанием частот их встречаемости. Можно задать диапазон частот (например, от 10 до 20 вхождений или ровно 15 вхождений) и сформировать список слов, количество употреблений которых лежит в границах этого диапазона. Если диапазон не задан, формируется полный частотный словарь текста.

Программа ЛЕКС2

Программа формирует список слов, обладающих указанными лексико-грамматическими характеристиками, например, находит все существительные, все причастия или все аббревиатуры, встретившиеся в тексте (области). Слова упорядочиваются по алфавиту, для каждого слова подсчитывается число его вхождений в исследуемый текст. Программа предназначена для анализа словарного состава текста.

Программа ЛЕКС3

Программа находит все вхождения в исследуемый текст (область) любых форм указанного (ключевого) слова и для каждого вхождения выдает контекст установленной длины - цепочку слов, находящихся от ключевого слова на расстоянии, не превышающем заданную длину. Программа удобна для анализа лексического состава текста и контроля используемых терминов и терминологических словосочетаний.

Программа ЛЕКС4

Программа находит в исследуемой области текста все слова, не входящие в формируемый в начале очередного сеанса словарь системы ЛИНАР, - т.е. слова, не знакомые очередному адресату. Для исправления текста следует либо заменить обнаруженные слова синонимами, либо расширить словарь системы. Возможно, что некоторые из обнаруженных слов являются известными системе словами, введенными с ошибками.

Программа ЛЕКС5

Программа осуществляет поиск каждой из обнаруживаемых в тексте (области) аббревиатур последовательно в трех списках:
N 3 – списке аббревиатур, вводимых непосредственно в тексте (этот список формируется динамически самой программой ЛЕКС5);
N 2 – списке, формируемом в начале работы с текстом на основе перечня используемых сокращений;
N 1 – словаре общепринятых сокращений.

В списке N 1 поиск ведется в последнюю очередь так как он, во-первых, самый большой, и во-вторых, если, например, в списках N 3 и N 1 присутствует одно и то же сокращение, но с различными расшифровками, то приоритет имеет сокращение из списка N 3. Результатом работы является список используемых в тексте аббревиатур с указанием их локализации в тексте.

Программа ЛЕКС6

Программа осуществляет контроль за переопределением известных системе аббревиатур. Если, например, в разделе 1.2. встретилась аббревиатура СВП (с расшифровкой в тексте – *«схема внешних прерываний»*), а в списке N 2 аббревиатура СВП сопоставлена термину *«субкомплекс внешней памяти»*, фиксируется ошибка: недопустимое переопределение аббревиатуры из перечня.

Программа ЛЕКС7

Программа проверяет правильность расшифровки, то есть тот факт, что аббревиатура читается в расшифровке по началам слов, причем некоторые слова расшифровки могут не участвовать в образовании аббревиатуры.

Пример работы программы:

```
Эта организация - центр переводов (ВЦП).
НЕСООТВЕТСТВИЕ АББРЕВИАТУРЫ И РАСШИФРОВКИ:
ВЦП - центр переводов
```

Программа ЛЕКС8

Программа ЛЕКС8 (без параметров) проверяет правильность оформления списка используемых в тексте аббревиатур (для отчета по НИР - это «Перечень условных обозначений, символов, единиц и терминов»). Предполагается, что каждая пара ***аббревиатура - расшифровка*** в перечне представлена одной строкой. В процессе обработки перечня заполняется список замечаний.

Пример работы программы:

```
ОБРАБАТЫВАЕТСЯ ПЕРЕЧЕНЬ АББРЕВИАТУР:
 БНК - бортовой нейрокомпьютер
 БНФ - бекусовская нормальная форма
 КПД - канал прямого доступа
 ОЗУ
 МПК - микропрограммируемый контроллер
 ОРЗ - общий регистр записи
 ПНП – перейти в неустойчивое положение
 СВП - субкомплекс внешней памяти
 СПТ - субкомплекс рабочего таймера

ЗАМЕЧАНИЯ:
4 : ОЗУ  * НЕТ РАСШИФРОВКИ
5 : МПК  * НАРУШЕНИЕ АЛФ. ПОРЯДКА
7 : ПНП  * РАСШИФРОВКА НЕ ЯВЛЯЕТСЯ ГРУППОЙ СУЩЕСТВИТЕЛЬНОГО
9 : СПТ  * НЕСООТВ: АББР.-РАСШ.
```

Стилистический контроль

Программы данного блока фиксируют внешние характеристики фраз, свидетельствующие о сложности их структуры, а следовательно, и о сложности восприятия смысла. Имеются, например, программы, контролирующие длину фраз, количество запятых, количество придаточных предложений, наличие во фразах текста длинных цепочек слов в родительном падеже (например, *значений аргументов программы пользователя*) или цепочек однокоренных слов (*пользователь может воспользоваться, транслятор транслирует*). Есть программы контроля стилистической окраски слов. В научно-технической литературе нежелательно употребление устаревших слов и канцеляризмов (*ибо; вышепоменованный*), жаргонизмов (*виндуза*), разговорных оборотов (*этот алгоритм, уж поверьте; . . .*). При обнаружении таких слов в тексте их рекомендуется убрать или заменить более нейтральными синонимами. Особый класс составляют слова, явно характеризующие специфику темы (предметной области), раскрывать которую иногда нежелательно. Например, в документе для внутреннего пользования можно употребить термин *военно-космический*, а в тексте сообщения, передаваемого по открытым каналам связи его целесообразно заменить (соответствующая программа предлагает слово-замену *специальный*).

Контроль структуры текста

Данные программы контролируют правильность оформления отдельных структурных частей текстового документа с точки зрения соответствующих нормативных требований (например, требований ГОСТа 7.32-81, регламентирующего правила оформления научно-технического отчета). Проверяется оформление титульного листа, списка исполнителей, реферата и других разделов документа.

Синтаксический контроль

Программа СИНТ1

Программа СИНТ1 находит в указанной области именные словосочетания вида <прилагательное> + <существительное> и <существительное> + <существительное в форме родит. падежа> и др. Программа может оказаться полезной при анализе лексического состава текста и при поиске терминологических словосочетаний, особенно в тех случаях, когда различные фрагменты текста написаны разными авторами (возможно, использующими близкие, но не совпадающие термины). Найденные программой словосочетания группируются вокруг «ключевого слова» - существительного, играющего роль синтаксической вершины словосочетания.

Ряд программ синтаксического контроля обнаруживает нарушения обычного (нейтрального) порядка слов и взаимного расположения групп слов. Такие нарушения могут затруднить восприятие текста.

Например: «*Раздел второй посвящен описанию новых алгоритмов».* или *«Использует этот алгоритм всего две вспомогательные переменные».*

Отметим, что иногда нарушение нейтрального порядка слов может намеренно использоваться автором текста с целью изменения логического ударения, усиления (*«Алгоритм этот очень эффективен!»*).

Программа СИНТ2

Программа СИНТ2 осуществляет контроль придаточных предложений с союзным словом *который*, а именно, проверяет однозначность установления связи между союзным словом и его словом-хозяином из главного предложения. В случае, когда таких слов-хозяев не обнаружено или их более одного, выдается соответствующая диагностика.

Пример работы программы:

Рассмотрим структуру памяти вычислительной машины, в которой хранятся команды.

СЛОВО которой ИМЕЕТ БОЛЕЕ ОДНОГО СЛОВА-ХОЗЯИНА В ГЛАВНОМ ПРЕДЛОЖЕНИИ: машины, памяти, структуру

Каждому каналу соответствует свое устройство, которые в свою очередь связаны с главной ЭВМ.

СЛОВО которые НЕ ИМЕЕТ СЛОВА-ХОЗЯИНА В ГЛАВНОМ ПРЕДЛОЖЕНИИ

Мощь языка Си - результат выявления его авторами потребностей программистов, которые возникают при программировании на языке ассемблера.

СЛОВО которые ИМЕЕТ БОЛЕЕ ОДНОГО СЛОВА-ХОЗЯИНА В ГЛАВНОМ ПРЕДЛОЖЕНИИ: программистов, потребностей, авторами

Пунктуационный контроль

Пунктуационные ошибки в реальных предложениях русского языка встречаются довольно часто.

Блок пунктуационного контроля системы ЛИНАР разработан на основе весьма полной пунктуационной модели русского языка. Полнота и корректность базовых знаний является основой достижения устойчивости и эффективности программных средств, реализованных на основе данной модели. В то же время блок пунктуационного контроля является «открытым», т.е. построен таким образом, чтобы обеспечить возможность работы средств адаптации и, при необходимости, введения новых правил пунктуации.

При проверке пунктуации можно использовать любое количество программ контроля, выбирая их при этом по различным признакам. Например, можно осуществлять проверку только тех правил, которые выявляют лишние знаки препинания, можно только тех, которые выявляют пропущенные знаки препинания и т.д. При подобной настройке может меняться совокупность пунктуационных правил, степень жесткости требований по соблюдению каких-либо условий и т. д., что позволяет оценивать качество текста с точки зрения различных категорий пользователей. Набор желаемых для данного сеанса модулей формируется в начале работы пользователем.

Пример работы программ пунктуационного контроля:

В ПРЕДЛОЖЕНИИ:
Только и развлечений , что кино раз в неделю
ЗАМЕЧЕНА ПУНКТУАЦИОННАЯ ОШИБКА.

В выделенном месте не должно быть данного знака препинания. В рассматриваемом случае запятая перед что не ставится .
Необходимо пояснение ошибки? (Д/Н)

Д

В безглагольном предложении перед союзом что в выражении только и … что , за которым следует имя существительное или местоимение, запятая не ставится.
Необходимы примеры правильного применения данного правила? (Д/Н)

Д

Только и денег что пятак в кармане.
Только и разговоров что о них двоих.

Семантический контроль

Программа СЕМ1

Программа обнаруживает несовпадение ожидаемых семантических признаков актантов (подлежащее, дополнения) глагола и признаков слов (групп слов), реально занимающих соответствующие позиции. Такое несовпадение мешает завершить анализ фразы, поскольку синтаксически допустимая связь не может быть установлена из-за семантических противоречий. Проверяя употребление в тексте глаголов, программа обращает внимание пользователя на «подозрительные» актантные конструкции.

Примеры работы программы:

Все рассматриваемые программы написаны на ассемблере.

НЕСОВПАДЕНИЕ СЕМАНТИЧЕСКИХ КЛАССОВ!
В ОПИСАНИИ ГЛАГОЛА "написать" СЕМ.-КЛАСС АКТАНТА:
=язык_программирования=
РЕАЛЬНЫЙ АКТАНТ ассемблере ИМЕЕТ СЕМ.-КЛАСС: =транслятор=

Схема прерываний подключается к магистрали.

НЕСОВПАДЕНИЕ СЕМАНТИЧЕСКИХ КЛАССОВ!
В ОПИСАНИИ ГЛАГОЛА "подключаться" СЕМ.-КЛАСС АКТАНТА:
=устройство=
РЕАЛЬНЫЙ АКТАНТ схема прерываний ИМЕЕТ СЕМ.-КЛАСС:
=структура2=

Программа СЕМ2

Программа проводит полный синтактико-семантический анализ фраз указанной области текста. При этом фиксируются случаи, когда фраза имеет (в контексте предметной области, к которой относится текст) более одной интерпретации, т.е. допускает неоднозначное толкование.

Пример работы программы:

Снижение напряжения вызвало отключение принтера.

НЕОДНОЗНАЧНАЯ ИНТЕРПРЕТАЦИЯ!
1 трактовка:
=причина= : снижение напряжения
=следствие= : отключение принтера
2 трактовка:
=причина= : отключение принтера
=следствие= : снижение напряжения

Программа CEM3

Программа CEM3 проверяет однозначность установления связи между личным местоимением и словом, на которое ссылается это местоимение. Если такое слово не найдено или же таких слов более одного, выдается соответствующая диагностика.

Пример работы программы:

Каждому каналу сопоставлено определенное устройство. Они, в свою очередь, связаны с главной ЭВМ.

ДЛЯ МЕСТОИМЕНИЯ они В ПРЕДШЕСТВУЮЩЕЙ ФРАЗЕ НЕ НАЙДЕНО СЛОВ, НА КОТОРЫЕ ЭТО МЕСТОИМЕНИЕ ССЫЛАЕТСЯ

Рассмотрим структуру памяти ЭВМ. Она состоит из двух основных частей.

ДЛЯ МЕСТОИМЕНИЯ она В ПРЕДШЕСТВУЮЩЕЙ ФРАЗЕ НАЙДЕНО БОЛЕЕ ОДНОГО СЛОВА,НА КОТОРОЕ ССЫЛАЕТСЯ ЭТО МЕСТОИМЕНИЕ:
ЭВМ, памяти, структуру

Программа CEM4

Программа проверяет, принадлежат ли значения количественно оцениваемых свойств описываемых в тексте объектов заданному диапазону. В случае, если значение свойства выходит за границы диапазона, процедура выдает соответствующую диагностику.

Пример работы программы:

Информация передается в сопроцессор АК-34 по 16 каналу.

ОБ'ЕКТ: сопроцессор АК-34 ГРУППА: 16 каналу
ВЫХОД ЗНАЧЕНИЯ ЗА ВЕРХНЮЮ ГРАНИЦУ ДИАПАЗОНА
(СОПРОЦЕССОР АК-34 ИМЕЕТ КАНАЛЫ: 0,1,2, ... 15)