Лингвистическое обеспечение АОТ-систем

Один из главных путей развития функциональных возможностей прикладных АОТ-систем и повышения качества их работы – создание и внедрение более полных и точных моделей естественных языков, более совершенных алгоритмов анализа и синтеза текста. Сегодня мы рассмотрим некоторые проблемы построения, формализации и компьютерной реализации моделей естественного языка на примере русской морфологии (словоизменения). Но перед этим введем одно важное понятие:

Лингвистические банки данных

Под ***лингвистическими банками данных*** (ЛБД) 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. 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. 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.

Отметим, что в качестве пользователя ЛБД может выступать как человек (исследователь-лингвист или разработчик программного продукта), так и тот или иной модуль компьютерной системы обработки текстов. В двух этих случаях требования к организации лингвистических банков данных и к степени эксплицитности, строгости и формальности представленных в них описаний естественного языка разнятся весьма существенно.

Ситуация здесь несимметричная. Пользователь-человек часто может извлечь нужную ему информацию из ЛБД, встроенного в компьютерную систему обработки текстов. Однако компьютерная система обычно не может извлечь нужную для ее работы информацию непосредственно из ЛБД, ориентированного на человека. Особенно остра эта проблема для флективных языков, в частности, для русского языка.

Так, во всех распространенных русскоязычных словарях (толковых, орфографических, словарях синонимов и антонимов и др.) входом в словарную статью служит так называемая начальная форма слова. Поскольку словари ориентированы на пользователя-человека, по умолчанию предполагается, что он знает правила русского словоизменения (склонения и спряжения) и может распознать в тексте любую форму интересующего его слова, т.е., восстановив начальную форму, добраться до соответствующей словарной статьи. Предполагается также, что он может решить и обратную задачу – употребить слово из словаря в требуемой грамматической форме.

При использовании словарей в составе компьютерных систем обработки текстов ситуация иная. Самоочевидные для человека грамматические свойства слова, определяющие особенности его склонения/спряжения, должны быть тем или иным способом явно представлены в компьютерном словаре и в программах морфологического анализа и синтеза, позволяющих определять грамматические признаки словоформ текста и генерировать слова в требуемой форме.

Как распределить знания о чрезвычайно сложных и запутанных правилах русского словоизменения между словарями и программными компонентами?

Здесь возможны два решения:

- в словаре описываются только словоизменительные признаки слов (тип и частные особенности склонения/спряжения), а работа по анализу и синтезу словоформ “поручается” программам морфологического компонента компьютерных систем;
- в словаре приводятся все формы слов, каждой из которых сопоставлены все необходимые признаки (в частности, грамматические: число, падеж, лицо, время, наклонение и др.).

В целом, задача построения и сопровождения лингвистически полного, обоснованного и покрывающего представительное подмножество выбранного естественного языка ЛБД, особенно в случае пользователя-программы, очень сложна. Ее решение требует привлечения квалифицированных специалистов в области лингвистики и инженерии знаний, создания необходимой инфраструктуры, серьезной финансовой и организационной поддержки (часто – на государственном уровне).

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 коммерческих автокорректоров (в том числе, ОРФО, Word), словари практически всех экспериментальных и коммерческих систем машинного перевода и других систем автоматической обработки текстов 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, например, этот 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 построении фонетического словаря.

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 начале 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 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 алгоритмов на 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 – образованных от другой основы – форм (*хорош-ий – лучше*) были организованы дополнительные входы в словарные статьи. Новое представление словаря трудно воспринимаемо для человека. Однако унификация и упрощение структур данных позволили значительно увеличить 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 (которое, по соглашению, 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 номер 7), адъективные (класс номер 8) и один класс спрягаемых (класс номер 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, форм единственного числа у слова *ножницы*, формы родительного падежа множественного числа у слова *мгла*) не мешает отнести его к М-классу номер 5.

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, однако, во избежание чрезмерного увеличения числа П-классов, в ФМРС используются другие методы.

Как исключения описываются случаи сочетания основы с «нестандартной» для данного П-класса и данной формы флексией: *-а* в форме именительного падежа множественного числа существительных вместо характерной для 1-го П-класса флексии *-ы* (*глаза*, но *заводы*), пустая флексия вместо флексии *-ов* в родительном падеже множественного 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 соответствующий 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.

Пример: `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 (*класс-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., муж.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) *квазибиологом*

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 СЛОВ1 словарные статьи кодируются по определенным стандартным правилам, в частности, заменяются символы, отсутствующие на клавиатуре (например, цифра в кружке заменяется на цифру в круглых скобках).

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 - цифре - определяется номер П-класса. Если второй элемент - не цифра (это означает, что слово изменяется по необычной модели), то СЛОВ1 фиксирует несовпадение номера С-класса с номером 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 автоматизацию не только этого, но и предыдущего этапа – этапа выявления незнакомых слов по характерным текстам.

Отдельные программы 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)

При пакетной обработке текстов на печать выдается так называемый «протокол формирования словаря», в который могут вставляться вопросы, адресуемые администратору. Рассмотрим фрагмент протокола диалога администратора-лаборанта с программой пословного анализа текста (будем считать, что слова: *колба, стержень, стекло, стечь* – отсутствуют в словаре):

```
 * РАБОТАЕТ ПРОГРАММА ФОРМИРОВАНИЯ СЛОВАРЯ ПО ТЕКСТУ *
УКАЖИТЕ, ПОЖАЛУЙСТА, ОТКУДА БУДЕТ ВВОДИТЬСЯ ТЕКСТ
(К - С КЛАВИАТУРЫ, <ИМЯ ФАЙЛА> - ИЗ ВНЕШНЕГО ФАЙЛА)
    К
ВВОДИТЕ ТЕКСТ. ПРИЗНАК КОНЦА - //.
```

```
    ВНУТРИ КОЛБЫ РАСПОЛАГАЕТСЯ СТЕРЖЕНЬ, ВЫПОЛНЕННЫЙ ИЗ
    СТЕКЛА. ЧЕРЕЗ СТЕРЖЕНЬ ... //
ВЫБЕРИТЕ РЕЖИМ РАБОТЫ, УКАЗАВ НОМЕР РЕЖИМА:
1 - ЗАПИСЬ СЛОВАРНЫХ СТАТЕЙ В БУФЕР БЕЗ ВЫДАЧИ ИНФОРМАЦИИ
2 - ЗАПИСЬ СЛОВАРНЫХ СТАТЕЙ В БУФЕР С ВЫДАЧЕЙ ИНФОРМАЦИИ
3 - ФОРМИРОВАНИЕ СЛОВАРНЫХ СТАТЕЙ С ПОМОЩЬЮ ПОЛЬЗОВАТЕЛЯ
    3
ПЕРВАЯ ФРАЗА: НЕЗНАКОМЫЕ СЛОВА: КОЛБЫ, СТЕРЖЕНЬ, СТЕКЛА
СЛОВО 'КОЛБЫ' - СУЩЕСТВИТЕЛЬНОЕ? (Д/Н)
    Д
К КАКОМУ РОДУ ОТНОСИТСЯ СУЩЕСТВИТЕЛЬНОЕ 'КОЛБЫ'? (М/Ж/С)
    Ж
ЯВЛЯЕТСЯ ЛИ ОНО ОДУШЕВЛЕННЫМ? (Д/Н)
    Н
К КАКОМУ РОДУ ОТНОСИТСЯ СУЩЕСТВИТЕЛЬНОЕ 'СТЕРЖЕНЬ'? (М/Ж/С)
    М
ЯВЛЯЕТСЯ ЛИ ОНО ОДУШЕВЛЕННЫМ? (Д/Н)
    Н
СЛОВО 'СТЕКЛА' - ГЛАГОЛЬНАЯ ФОРМА? (Д/Н)
    Н
К КАКОМУ РОДУ ОТНОСИТСЯ СУЩЕСТВИТЕЛЬНОЕ 'СТЕКЛА'? (М/Ж/С)
    С
СФОРМИРОВАНЫ СЛОВАРНЫЕ СТАТЬИ С ОСНОВАМИ:
КОЛБ-, СТЕРЖН-, СТЕКЛ-.
ВТОРАЯ ФРАЗА: ...
 ...
```

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

Рассмотрим примеры, показывающие возможность комбинирования отдельных программ библиотеки «Русская морфология». 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
```