A. Gelbukh. *An effectively implementable model of morphology of an inflective language* (in Russian). J. Nauchno-Tehnicheskaya Informaciya (NTI), ISSN 0548-0027, ser. 2, vol. 1, Moscow, Russia, 1992, pp. 24–31.

ЭФФЕКТИВНО РЕАЛИЗУЕМАЯ МОДЕЛЬ МОРФОЛОГИИ ФЛЕКТИВНОГО ЕСТЕСТВЕННОГО ЯЗЫКА

А.Ф.Гельбух

Излагается структура описания морфологического строения флективного языка, обеспечивающая экономичную по быстродействию и памяти программную реализацию на ПЭВМ алгоритмов анализа синтеза, нормализации форм и исправления ошибок. Рассматривается полный и точный словарный анализ и синтез. Описывается только суффиксальное словоизменение, чередования сводятся к супплетивизму и дополнительным суффиксам. Примеры приводятся для русского языка.

ВВЕДЕНИЕ

Под задачей морфологического анализа понимается сопоставление словоформе набора значений ее морфологических характеристик (таких, как часть речи, падеж, число и др.) и имени лексемы в виде ее словарного номера или нормализованной формы, например:

красными → лексема = КРАСНЫЙ,
часть речи = прилагательное,
падеж = творительный,
число = множественное,

либо установление того, что сведений об этой форме в системе нет. Под задачей морфологического синтеза понимается обратное сопоставление. Как анализ, так и синтез могут быть неоднозначными: например, *стекла* — существительное СТЕКЛО либо глагол СТЕЧЬ; творительный падеж множественного числа от ДВЕРЬ — *дверями* либо *дверьми*.

Анализ (синтез) является полным, если учитываются все возможные варианты разбора (синтеза) словоформы. Анализ является точным (в отличие от приближенного), если выдаются все характеристики словоформы, не порождаются «лишние» омонимы, и выводы системы не носят эвристического характера, а основыва ются только на словарных данных.

Под задачей нормализации (лемматизации) понимается сопоставление словоформе либо номера лексемы (общего для словоформ одной лексемы и разных для словоформ разных лексем), либо ее так называемой словарной формы (например, *иду*, *шел*, *шедшими* — ИДТИ; *ножницами*, *ножниц* — НОЖНИЦЫ; *рада*, *рады* — РАД). В первом случае мы будем рассматривать задачу нормализации как часть задачи анализа, а во втором — еще и как часть задачи синтеза. Естественно, при омонимии анализируемой словоформы нормализация может быть неоднозначной.

В последние десятилетия многими исследователями были разработаны системы морфологического анализа, синтеза, нормализации и/или исправления ошибок, как для русского языка [1–5, 11], так и для многих других языков мира. Однако до сих пор в научном обиходе нет эффективной, доступной, универсальной морфологической системы с достаточно полным отделением лингвистической информации от программной. Автором предпринята попытка восполнить этот пробел, реализовав в виде библиотеки объектных модулей систему с такими свойствами. В настоящей работе излагается используемая в нашей реализации структура представления лингвистической информации, описывающей конкретный естественный язык, имеющая, как нам кажется, и самостоятельный интерес.

Согласно сформулированному Ингве принципу отделения лингвистической информации от программной, при разработке системы должны создаваться отдельно программное обеспечение, полностью настраиваемое на заданный вариант описания языка или даже на другой язык (в некотором классе), и лингвистическое обеспечение, заданное так, чтобы использующие его программы можно было перерабатывать и совершенствовать независимо. Общей точкой для этих двух компонентов является модель естественного языка, задающая общую структуру его описания. Такая модель должна оперировать понятиями, достаточно удобными для программной реализации, и в то же время предоставлять достаточно мощные выразительные возможности лингвисту.

Приводимое в настоящей работе описание лингвистической модели можно рассматривать и как описание общих принципов построения программного обеспечения системы, удовлетворяющей указанным выше требованиям, и как описание структуры ее лингвистического обеспечения.

Для такого инструмента, как универсальная морфологическая система, важным свойством является языковая независимость в некотором классе языков. Различные варианты лингвистического обеспечения для реализованной автором программы были составлены в первую очередь для русского языка. Однако есть основания полагать, что описания с единых позиций в рамках предлагаемой модели могут быть составлены для широкого класса языков с суффиксальным словоизменением.

Существенными требованиями к практически работающей системе, кроме полноты и точности, в настоящей работе полагаются большой объем словаря, высокое быстродействие и небольшая потребность в оперативной памяти. Программы системы должны исполняться на современных малых ЭВМ, и в первую очередь на распространенных типах ПЭВМ.

Реализация прикладных систем на малых машинах и ПЭВМ накладывает особенности на всю архитектуру системы и модель предметной области, служащую для представления знаний в системе. Предполагается, что машина, на которой должна реализовываться предлагаемая лингвистическая модель, обладает следующими особенностями, свойственными современным малым машинам.

Первая: обьем оперативной памяти недостаточен для хранения всего словаря. Под это условие подпадают машины (и операционные системы) с эффективно доступной одному процессу памятью до одного мегабайта. Мы предполагаем, что результаты работы программы должны допускать использование в системах смысловой обработки информации, а смысл обычно привязывается к лексемам. Поэтому номера лексем не должны изменяться при пополнении словаря. Это требование обычно приводит к необходимости хранения в словаре большого количества таких номеров. Уже поэтому крупный словарь, позволяющий определять номера лексем, трудно сделать по размеру существенно меньше одного мегабайта.

Вторая особенность состоит в том, что за один акт прямого доступа к внешней памяти на системно-аппаратном уровне может быть считан лишь фрагмент фиксированного размера порядка одного килобайта - блок файла. При этом любой разумный объем обработки данных в памяти без больших переборов выполняется быстрее, чем считывание произвольного блока данных с внешнего носителя. Поскольку последнее утверждение справедливо и для имеющихся в настоящее время плат расширения памяти, то такая память также должна рассматриваться как внешняя. Поэтому наши рассуждения прямо переносятся и на этот случай.

При реализации небольшой объем требуемой памяти обеспечивается небольшим размером таблиц, допускающих «вычисление» большого количества сложных окончаний

(например, *-а-ющ-ими-ся*) путем комбинации простых, а также хранением словаря на внешнем носителе. Быстродействие при этом обеспечивается (в режиме анализа и исправления ошибок) специальной структурой словаря, подробно рассмотренной нами в [6].

При разработке модели было уделено внимание проблеме понимания ошибочно написанных словоформ и выдаче разумных гипотез об их исправлении. Эффективно распознается один класс характерных ошибок: ошибки, происходящие от выбора для слова основы, суффикса или формы, правильной для других слов языка, но не для данного слова. Такие ошибки могут допускать лица, не владеющие активно языком. Алгоритму исправления ошибок иного вида — опечаток — посвящена другая наша работа [7].

Выбор терминологии представлял большую трудность для автора, поскольку в лингвистике не существует общепринятых терминов для обозначения необходимых нам конструкций [8]. Поэтому в случаях употребления нами терминов в значении, отличном от общепринятого, все приводимые термины следует при чтении данной работы формально понимать так, как они определяются нами.

ОСНОВНАЯ СХЕМА МОДЕЛИ

Рассматривается флективный (или, как частный случай, агглютинативный) естественный язык, в котором, однако, из возможных типов аффиксов рассматриваются только суффиксы. Таким образом, лексема характеризуется цепочкой букв — основой (основы лексемы могут быть разные для разных ее форм, что называется супплетивизмом основ), а морфологические характеристики конкретной словоформы определяются присоединенными справа к основе одной или несколькими буквенными цепочками — суффиксами. В нашей модели не рассматриваются чередования букв. Они сводятся к супплетивизму основ или аффиксов либо к более мелкому дроблению словоформы на аффиксы.

Основная схема построения модели состоит в следующем. Словоформа представляется расчлененной на определенное для данной части речи число частей, которые будем называть позициями. Все объекты рассмотрения и свойства будем далее привязывать к позициям или к связям между различными позициями. Позиции пронумеруем начиная с нуля.

Буквенные цепочки, стоящие на различных позициях словоформы, будем называть морфами. Например, в словоформе *читающийся* выделяются четыре позиции: *чита-ющ-ий-ся*. На нулевой позиции стоит морф *чита-*, на первой — морф *-ющ-*, на второй *-ий-* и т.д. В словоформе же *читающимися* на второй позиции стоит морф *-ими-*. Морф, стоящий на нулевой позиции, мы называем основой.

С каждым морфом связывается некоторое смысловое значение, например, сочетание падежа, числа и т.п. Сочетание значений морфологических характеристик, кумулятивно выражаемое одним морфом, будем называть граммatемой (ср. [8, с. 552, статья "Флективность", п. 1]). Объединение грамматем, соответствующих всем морфам словоформы, образует полный набор значений характеристик словоформы: так, соответствие

чита-ющ-ий-ся ↔ лексема = ЧИТАТЬ,
чр = глагол,
залог = действ,
вр = наст,
род = муж,
чсл = ед,
возвратность = возвр

получается из соответствий

чита- ↔ лексема = ЧИТАТЬ, чр = глагол,
-ющ- ↔ залог = действ, вр = наст,
-ий- ↔ род = муж, чсл = ед,
-ся ↔ возвратность = возвр

Здесь и далее используются сокращенные названия грамматических категорий и их значений.

В лингвистических таблицах системы декларативно представлена информация трех типов, одного — для связи между грамматемами и морфами, стоящими на одной позиции, и двух — для связи между разными позициями словоформы, а именно:

1) Морфы соответствуют грамматемам. Так, в именительном падеже единственного числа надо писать *-ий-*, а в родительном *-его-*.
2) Морфы управляют морфами. Так, после *-ющ-* пишется *-ий-*, но после *-ем-* пишется *-ый-*. Это синтактика морфов одной позиции по отношению к морфам другой.
3) Грамматемы управляют грамматемами. Так, можно писать *-ющ-*(*ий*)*-ся* и нельзя *-ем-*(*ый*)*-ся*, поскольку возвратность не совместима со страдательным залогом. Это синтактика грамматем одной позиции по отношению к грамматемам другой.

Вся информация в нашей системе делится на две части. Словарная информация приписывается основам и характеризует лексику языка. Иерархически организованная информация, приписанная морфам и грамматемам, характеризует грамматику языка независимо от лексики.

Существенно, что среди правил типов 2 и 3 фигурируют только правила вида «после того-то можно/нельзя/нужно писать то-то», но не рассматриваются правила вида «перед тем-то пишется то-то». Такое ограничение, как правило, не обременительное для лингвиста, упрощает реализацию, делая проход по слову однонаправленным.

Дальнейшее изложение построим по возрастанию сложности. Поскольку основная сложность русской морфологии заключается в описании глаголов, примеры на основе глаголов были бы переусложненными, и мы ограничимся, в основном, примерами на основе существительных и прилагательных. Однако возможности данной модели в случае русского языка полностью используются лишь при описании глаголов и причастий.

ТАБЛИЦА МОРФОВ

В общих чертах структуру представления лингвистической информации в модели можно проиллюстрировать общепринятой [1, 3] схемой рис. 1.

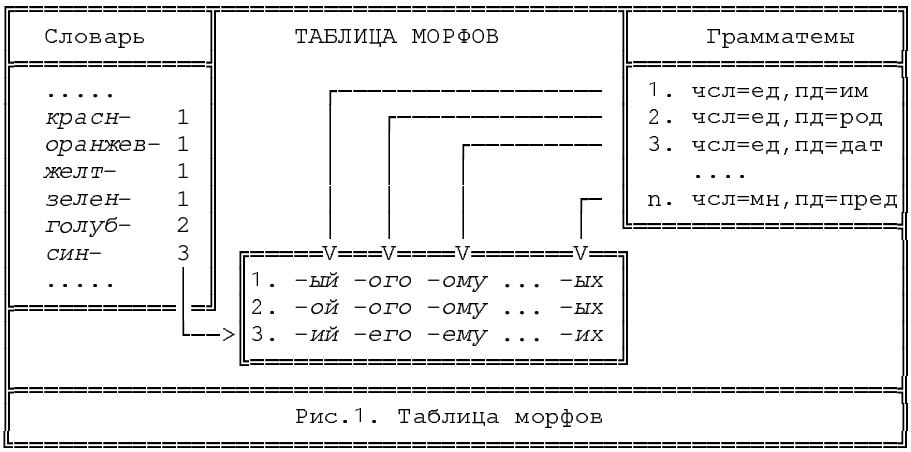

Рис.1. Таблица морфов

Алгоритмы как анализа, так и синтеза очевидны. При анализе следует рассмотреть все возможные варианты составления входного слова из левой части, имеющейся в словаре, и правой части, имеющейся среди морфов. Порядок и способ перебора выбирается программистом. Наша модель построена так, чтобы такой перебор был вычислительно эффективным [6]. Аналогично все возможные способы следует использовать при синтезе.

Таблицей морфов мы будем называть прямоугольную матрицу, составленную из морфов, столбцы которой (представляющие собой формальные морфемы, ср. [8, стр. 312]) занумерованы всеми возможными для лексем, имеющихся в словаре, грамматемами (для данной позиции, см. п. «Случай нескольких суффиксов»), а строки — различными вариантами правописания слов по данной позиции (формальными парадигматическими классами).

ЧЕРЕДОВАНИЕ ОСНОВ И МАСКИ

Для отражения явления супплетивизма, как истинного, так и проистекающего из чередований букв, некоторые основы из хранящихся в словаре обьединяются в группы, представляющие одну и ту же лексему. Такие основы мы назовем чередующимися. Например, лексеме РЕБЕНОК соответствуют чередующиеся основы *ребенок*-(*), *дет*-(*и*) (истинный супплетивизм) и *ребенк*-(*а*) (чередование, формально описываемое как супплетивизм). Каждая из этих основ имеет ущербную парадигму, то есть может сочетаться не со всеми грамматемами — так, для основы *дет*- запрещены все формы единственного числа.

Для описания ущербности парадигм введем понятие маски [3, 9] как строки знаков, длина которой равна количеству столбцов таблицы морфов, т.е. количеству грамматем данной позиции данной части речи. Каждый из этих знаков задает информацию о возможности для основы иметь данную грамматему в составе набора ее характеристик, точнее — сведения о возможности выбрать данную морфему (столбец таблицы морфов).

На рис. 2 дан пример. Номера масок приписаны основам после номеров строк таблицы морфов через косую черту. В словаре чередующиеся основы одной лексемы показаны соединенными. Символы в масках имеют следующий смысл: «+» — разрешено, «–» — запрещено. Знаком * в таблице морфов обозначен пустой морф. Видно, что основам слов танк и тать разрешены все формы («+» в маске), а основам *ребенок*-, *ребенк*- и *дет*- знак «–» маски исключает формы **ребенока*, **ребенки* и **деть*. Указано, что основы *ребенок*-,

ребенк- и *дет-* представляют одну лексему, то есть им должен соответствовать один и тот же идентификатор.

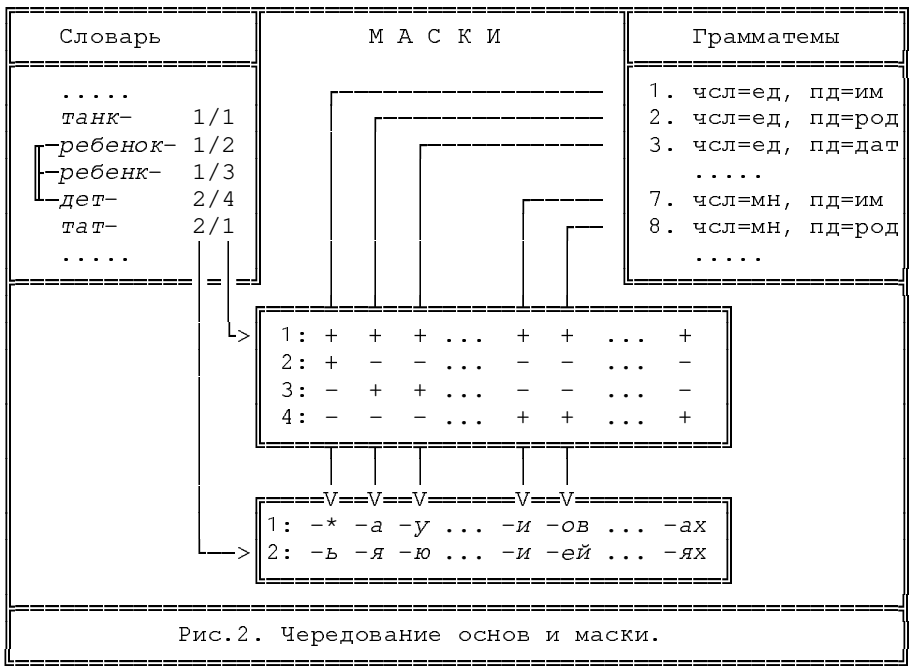

Рис.2. Чередование основ и маски.

Кроме информации о разрешении/запрещении какой-либо грамматемы для основы, маски можно использовать для сообщения более подробной информации. Так, в выполненной автором реализации используются знаки «%» и «&», выражающие соответственно затрудненность и факультативность образования формы.

Случаи чередований могут формально обслуживаться не только механизмом описания супплетивизма и масками, но и добавлением новых (лингвистически не содержательных) позиций так, например, беглое *-о-* в словах ребенок, станок можно было бы формально описать как форматив *-ок-/-к-*, занимающий отдельную позицию. Кроме того, можно было бы, конечно, завести для таких слов отдельную строку таблицы морфов вида *-ок*, *-ка*, ..., *-ках*.

Фактическое отсутствие в алгоритме понятия чередований в основе существенно облегчает и ускоряет анализ (но не синтез). Опыт реализации показывает, что рост объема словаря при увеличении за счет этого числа основ можно считать приемлемым (благодаря применению в словаре сжатия по Куперу [6]).

Итак, механизм супплетивизма и масок описывает два явления — собственно супплетивизм и чередования. Третья функция механизма масок состоит в описании собственно ущербных парадигм — невозможности образования некоторых форм для отдельных лексем или для целых классов (**деревенск*, **радый*). Четвертая функция механизма супплетивизма/масок — описание вариативности и факультативности (*священн* / *священен*, *псалтирем* / *псалтирью*). При этом области в масках различных основ одной лексемы, помеченные знаком «+», просто перекрываются.

СЛУЧАЙ НЕСКОЛЬКИХ СУФФИКСОВ

Рассмотрим теперь слова, делящиеся на несколько позиций. В нашей модели принимается, что для одной (формальной) части речи каждая словоформа формально делится ровно на *N* позиций, возможно, выраженных пустыми морфами, причем *N* не зависит от словоформы. Снова будем нумеровать позиции начиная с нуля, где нулевая позиция есть основа.

Подчеркнем, что некоторые позиции в словоформе могут быть фиктивными и вводятся лишь для единообразия описания. Например, словоформы лексемы ДЕЛАТЬ *дела-ющ-ими-ся* и *де-ла-ю-#-** формально считаются состоящими обе из четырех позиций, причем позиция со знаком # фиктивна, а со знаком * значима и заполнена пустым морфом (ср. *дела-ю-#-сь*).

В предыдущем изложениии мы игнорировали тот факт, что приведенные на рис. 1 и 2 прилагательные на самом деле имеют не две (включая нулевую), а три позиции: *важн-ейш-ий*. Для учета нескольких позиций строятся все упоминавшиеся таблицы для каждой позиции в отдельности, кроме нулевой позиции (список форм для нулевой позиции — основы — обсуждается ниже).

Пример таблиц для первой позиции прилагательных приведен на рис. 3. Символом * обозначена пустая грамматема. В словаре все данные указываются для каждой позиции по отдельности (ср. рис. 2).

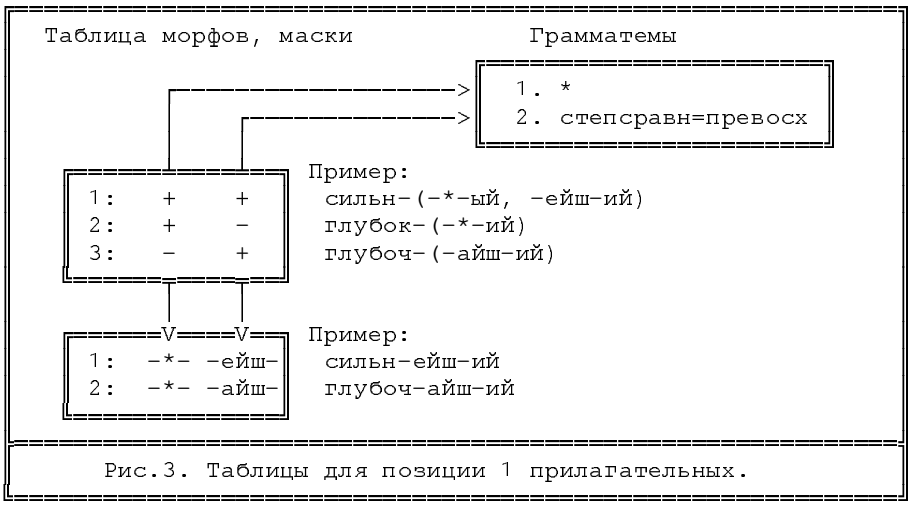

Рис.3. Таблицы для позиции 1 прилагательных.

Для описания лексем, присоединяющих несколько суффиксов, вводятся также обсуждаемые ниже механизмы взаимной зависимости между морфами и между морфемами разных позиций одной словоформы.

СИНТАКТИКА МОРФОВ

Для каждой позиции словоформы все предыдущие позиции в нашей модели формально играют роль, подобную роли основы, а все последующие подчиняются ей аналогично тому, как в простейшем случае выбор морфа подчинен основе. Так, всегда пишется ...*-ейш-ий*, но не ...*-ейш-ый*. В соответствии с этим составляются как бы мини-словари морфов для каждой позиции. В них для морфов указываются номера строк таблицы морфов всех следующих позиций подобно тому, как это указывается в словаре основ (ср. словарь на рис. 1).

При этом допустимо специальное значение «*», означающее отсутствие указания. В этом случае используется значение, указанное при морфе предыдущей позиции, и т.д. вплоть до суффикса 0 — основы, если это необходимо.

Таким образом, в дополнение к показанным на рис. 3 таблицам требуется еще одна, которую назовем списком морфов. Для примера рассмотрим список морфов позиции 1 части речи прилагательных. Заметим, что на рис. 1 фактически приведены таблицы для позиции 2: *красн-*-ый*, *красн-ейш-ий*. Как и в словаре на рис. 1, припишем каждому морфу позиции 1 номер строки таблицы морфов следующей позиции (*-ый/-ий*), используя номера строк для позиции 2 с рис. 1 и морфы позиции 1 с рис. 3:

-айш-	3	пример: *ближ-айш-ий*	
-ейш-	3	*сильн-ейш-ий*	(1)
-*-	*	*син-*-ий*, *красн-*-ый*	

С учетом позиции 1 (*-ейш-/-айш-*) основам прилагательных в словаре на рис. 1 будет приписан не один номер строки, а два — первый для *-айш-/-ейш-*, а второй для *-ый/-ой/-ий*.

Рассмотрим пример работы алгоритма синтеза. Номера строк и столбцов относятся к таблице морфов, причем ее столбцы соответствуют грамматемам. Используются таблицы с рис. 1 и 3 и список морфов (1). Показан процесс синтеза, при анализе следует действовать в обратном порядке:

```
На входе : КРАСНЫЙ (т.е. номер NN), степсравн=*, пд=дат.
В словаре: красн-, 1-я позиция: 1/1, 2-я позиция: 1/1.
Позиция 0: красн-
Позиция 1: строка 1 (от позиции 0), столбец 1: -*-
Позиция 2: строка 1 (от позиции 0), столбец 3: -ому
На выходе: красному

На входе : КРАСНЫЙ, степсравн=превосх, пд=дат.
В словаре: красн-, 1-я позиция: 1/1, 2-я позиция: 1/1.
Позиция 0: красн-
Позиция 1: строка 1 (от позиции 0), столбец 2: -ейш-
Позиция 2: строка 3 (от позиции 1), столбец 3: -ему
На выходе: краснейшему
```

Впервом случае во второй позиции выбирается строка 1, поскольку при выбранном морфе -*- из первой позиции в списке морфов (1) стоит пустое указание «*». Тогда просматривается предыдущая позиция, в данном случае нулевая — словарь, и, поскольку там для второй позиции стоит номер 1, он становится действующим.

Во втором случае для второй позиции выбирается строка 3, поскольку именно она указана в списке морфов (1) при *-ейш-*. Номер строки из словаря игнорируется. Конечно, в процессе синтеза необходимо было убедиться, что во всех нужных масках в соответствующих столбцах стоят знаки «+». Кроме того, следовало просмотреть все чередующиеся основы указанной лексемы.

Точно так же действует механизм управления морфами в случае большего числа позиций. В этом случае в списке морфов i-го суффикса (в словаре при $i = 0$) при каждом морфе указываются номера строк для последующих суффиксов, и действующей для каждого суффикса будет строка таблицы морфов, выбираемая для него самым ближним левым суффиксом, у которого это значение отлично от «*».

Дополнительной функцией механизма управления морфов является организация «древесной» структуры полного набора словоформ лексемы. Для этого, как указывалось,

некоторые позиции в словоформах следует описать как фиктивные. С этой целью допускается вместо номера строки при морфе в списке морфов указывать, со специальной пометой, сразу номер столбца таблицы морфов, то есть сразу необходимую грамматему. Соответствующая позиция, в случае выбора для нее по общему правилу именно такого специального «номера строки», считается фиктивной, однако выбранная грамматема добавляется к общему набору характеристик.

Например, если морф *-ющ-* глаголов в списке морфов дает ссылку на строку таблицы морфов следующей позиции, содержащую *-ий-*, а морф *-ть-* ссылку на столбец, которому соответствует, скажем, пустая грамматема, то при анализе словоформы *чита-ющ-ий-ся* будут просмотрены все четыре позиции, а при анализе слева направо ([6], см. тж. п. «Реализация») словоформы *чита-ть-#-ся* после первой будет просматриваться сразу третья позиция. Это приведет как к некоторой экономии времени, так и к отсутствию в выдаче морфного разбора не значимых пустых позиций.

СИНТАКТИКА МОРФЕМ

Ключевое значение в наших рассмотрениях имеет номер формальной морфемы. Это номер столбца таблицы морфов, он же номер элемента (столбца) маски или номер набора характеристик грамматемы, приписываемого словоформе при выборе данного столбца. Например, форме *красн-ейш-ий* соответствует морфема номер 2 позиции 1 (*-ейш-*) и морфема номер 1 позиции 2 (*-ий*) (см. рис. 1 и 3).

Более точно, следует говорить о том, что данному номеру морфемы однозначно соответствуют как столбцы таблицы морфов и списка масок, так и некоторая грамматема, поскольку грамматемы, приписанные различным морфемам, могут, вообще говоря, совпадать. Таким образом, набору (вектору) морфем однозначно соответствует словоформа и набор ее характеристик. Обратное соответствие номера морфемы грамматеме может и не быть однозначным.

Однако там, где это не вызывает недоразумений, мы упоминаем морфемы не по номерам, а по характерному морфу, знаку маски или грамматеме и говорим, допуская вольность речи, о морфеме *-ейш-* или морфеме {степсравн = превосх}, имея в виду морфему номер 2 позиции 1 прилагательных.

Взаимосвязь морфем различных позиций может заключаться в несовместимости (невозможности употребления в одной словоформе) или в затрудненности совмещения этих морфем. Так, основа *чита-* может иметь как отдельно морфему {залог = страд, время = прош}, соответствующую *-нн-* (*чита-нн-ый-**), так и отдельно форму {возвр} (*чита-ющ-ий-ся*), однако ни одна словоформа не может иметь сочетание морфем **(чита)-нн-ый-ся*.

В естественном языке удается так подобрать набор грамматем, что подобные запреты, во-первых, являются свойствами морфем, а не индивидуальными свойствами слов (так, *...-ем-ый-ся* запрещено для всех слов), и, во-вторых, касаются только пар морфем, а не более сложных их сочетаний. В случае невыполнения последнего условия можно разбить одну морфему на несколько по-разному сочетающихся с другими, причем в естественном языке это часто хорошо соответствует лингвистической неоднозначности (омосемии) морфов. Таким образом, каждой морфеме (столбцу таблицы морфов) можно сопоставить по одной маске на морфемы последующих суффиксов.

Кроме обязательных, грамматических запретов такие маски могут содержать и обязательное разрешение. Так, *-айш-* всегда разрешает падежные окончания, даже в том случае, если по словарю они запрещены основе. Например, при основе *глубоч-* в словаре

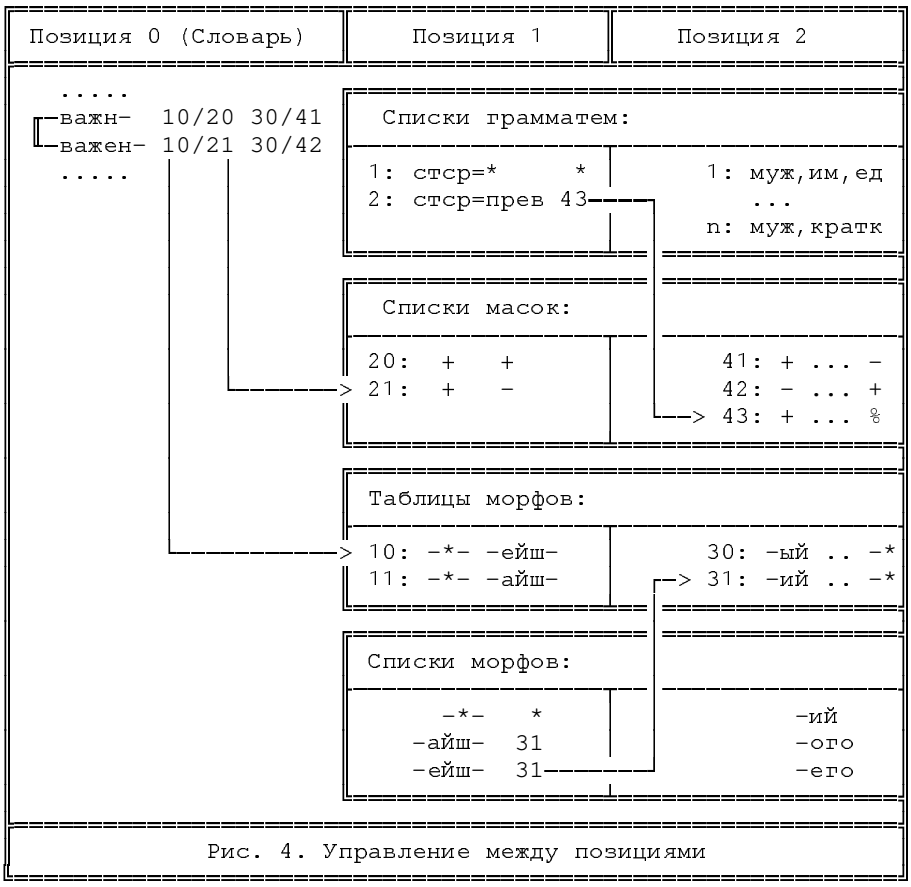

Рис. 4. Управление между позициями

должна быть указана маска, запрещающая *-ий*, поскольку пустой морф второй позиции -*- не может дать такого запрета. Однако *-айш-* разрешает его: *глубоч-айш-ий.*

Как и в случае управления между морфами, вводится специальное значение «*» элемента маски, означающее отсутствие управляющей информации. Таким образом, в списке масок возможны, например, маски, содержащие «–» для запрета некоторых форм и «*» для отсутствия запрета или разрешения других форм. Как и раньше, значение «*» игнорируется при нахождении ближайшего управляющего значения: действующим считается ближайшее (слева) значение, отличное от «*». Такое соглашение, возможно, соответствует тенденции естественного языка ставить текстуально рядом семантически связанные суффиксы. Номер маски «***...*», состоящей из всех «*», для сокращения обозначается «*». Он игнорируется при поиске управляющего номера маски.

На рис. 4 дан пример. Список грамматем рис. 3 здесь дополнен номерами управляющих масок для позиции 2 (в словаре номера строк и масок для одной позиции даны через косую черту; все номера для большей наглядности сделаны различными). По рис. 4 легко проверяется возможность построения словоформы *важн-*-ый*: морфемой 1 (*-ый*) позиции 2 управляет с помощью маски морфема 0 — словарь, поскольку морфема 1 (-*-) позиции 1 содержит в поле номера управляющей маски (в списке грамматем) пустое значение «*»; далее, возможность построения затрудненной краткой словоформы *важн-ейш-** (мы оставляем за пределами обсуждения лингвистическую спорность данного конкретного примера): морфемой *n* позиции 2 управляет морфема 2 позиции 1, выбирающая для нее маску 43; и, в то же время, невозможность совмещения двух морфем — построения

словоформы **важн*-*-*, поскольку в этом случае морфемой n позиции 1 управляет маска 41, полученная из словаря, которая содержит в данной позиции «–».

Дополнительно вводится такая же, как в случае управления между морфами, возможность указать фиктивность некоторых из последующих позиций. Для этого следует поместить в поле номера управляющей маски со специальной пометой сразу номер морфемы (см. конец предыдущего пункта). Так, например, грамматема {инфинитив} может блокировать поиск падежных окончаний причастий (ср. там же).

Изложенный механизм, вряд ли оправданный при описании русских прилагательных, как в примере рис. 4, играет, однако, важную роль, например, при описании русских причастий.

СПИСКИ ГРАММАТЕМ

Рассмотрим подробнее структуру списка грамматем. Заметим, что в языке могут существовать «омосемичные» ряды морфов, во всех словоформах порождающие омонимию наборов характеристик; другими словами, «омонимичные» грамматемы, в любой словоформе дающие одинаковые морфы. Таковы, например, для русских существительных наборы «пд=им, род=муж, чсл=ед, од=неодуш» (*стоит станок*) и «пд=вин, род=муж, чсл=ед, од=неодуш» (*вижу станок*), для прилагательных — наборы «род=жен, чсл=ед, пд=род/дат/твор/предл» (*нет белой бумаги*, *к белой бумаге*, *белой бумагой*, *о белой бумаге*). В нашей модели такое явление выражается в совпадении столбцов таблицы морфов.

В этом случае нет смысла различать эти столбцы до тех пор, пока речь не идет собственно о грамматемах. Поэтому допускается возможность поставить несколько грамматем в соответствие одному столбцу таблицы начертаний (морфеме). Окончательно список грамматем (ср. рис. 4) имеет следующий вид. Элемент этого списка (для каждой позиции, кроме нулевой, элемент соответствует номеру столбца таблицы морфов) включает, во-первых, одну или несколько грамматем - наборов значений характеристик, и, во-вторых, перечисление номеров масок на каждую из следующих позиций. Впрочем, большинство из этих номеров на практике оказывается равными «*».

ЧАСТИ РЕЧИ. СЛОВАРЬ ОСНОВ

В предыдущем изложении использовались два варианта таблиц — рис. 2 (существительные), с одной стороны, и рис. 4 (прилагательные), с другой. При этом рассматривались одинаково устроенные таблицы и алгоритмы, различающиеся количеством суффиксов и конкретным заполнением таблиц. Такие варианты называются в модели техническими (формальными) частями речи. Таким образом, строится несколько различных реализаций лингвистической модели, как если бы имелось несколько разных языков (что, таким образом, также допустимо в нашей модели).

Как правило, технические части речи хорошо соответствуют известным грамматическим частям речи, однако это соответствие не всегда удобно сохранять. Так, возможно, что лингвист пожелает отнести существительные адъективного склонения к одной технической части речи с прилагательными; слова типа *учащийся* — с глаголами (причастиями); все неизменяемые слова — наречия, предлоги, союзы — удобно отнести к одной технической части речи. Собственно грамматическая часть речи является при этом словарной информацией. Таким образом, словарь играет в некотором смысле роль таблицы морфов и списка морфов для позиции 0.

Единственная таблица, релевантная для позиции 0 — список грамматем, строящийся так же, как для остальных позиций. В грамматемы позиции 0 входят характеристики, привносимые в набор характеристик словоформы самой основой, в частности, грамматическая часть речи. Ссылки на данную таблицу помещаются в словаре основ. Именно так основа слова булочная, отнесенная к технической части речи «прилагательные», получит в списке характеристик граммему {часть речи=сущ}, а две разные основы слова выхухоль получат соответственно граммемы {род=муж} и {род=жен} (вариативно нет *выхухоля* / *выхухоли*). У большинства «нормальных» лексем грамматемы позиции 0 у разных основ совпадают.

Имеющиеся в списке грамматем 0 номера масок на следующие позиции задают обязательные (грамматические, не зависящие от лексемы) законы сочетаемости, которые, таким образом, можно не указывать в словаре. Например, пусть в таблице морфов существительных имеется отдельный столбец для описания дополнительных форм творительного падежа женского рода (*женщин-ой* и *женщин-ою*, но только *инженер-ом*). Тогда существительным мужского рода эта морфема запрещается сразу грамматемой из списка 0. Естественно описываются и исключения: слово *мужчина* (*мужчиною*) ссылается на другую морфему позиции 0, отличающуюся от морфемы, на которую ссылается слово инженер, только номерами масок, но не набором характеристик.

В дополнение к грамматическим запретам сочетаемости в словаре могут указываться и индивидуальные, лексические запреты для конкретной основы конкретной лексемы (см. п. Синтактика морфем). Это необходимо в случае чередования основ, а также для описания исключений, однако при этом не нужно дублировать сложную систему грамматических запретов.

Итак, в словаре основ указывается следующая информация. Во-первых, сама основа. Во-вторых, идентификатор лексемы. В-третьих, техническая часть речи — номер набора таблиц. В-четвертых, номера строк таблиц морфов и номера масок для каждой позиции. При этом вместо номера строки и/или номера маски может быть указан сразу номер морфемы (элемента списка грамматем). Для нулевой позиции это является обязательным требованием. Для ненулевых позиций это поле часто бывает заполнено значением «*».

И, наконец, по словарю основ должен обеспечиваться механизм перехода от одной основы лексемы к другой. При этом для целей синтеза и исправления ошибок желательно, чтобы получение всех масок различных основ конкретной лексемы было как можно более легким. Поэтому в нашей реализации для наиболее продуктивных классов номера масок технически хранятся отдельно от словаря — в оперативной памяти, а в словаре указывается общий для всех основ одной лексемы номер такой группы масок. Кроме того, механизм перехода между основами реализован путем особого назначения идентификаторов лексем и не требует дополнительных ссылок. Последние два приема сообщены автору Г. Б. Савиной [9].

РЕАЛИЗАЦИЯ. АНАЛИЗ СЛОЖНЫХ СЛОВ

В выполненной реализации лингвистического обеспечения для русского языка был использован словарь А. А. Зализняка [10] и словарь тезауруса ВНТИЦентра ГКНТ СССР, преобразованные в словари словоизменительных основ.

Алгоритм текстуального разбора построен «слева направо»: сначала по словарю определяется основа (это происходит за одно обращение к дисковой памяти на слово; подробности см. [6]), затем устанавливается техническая часть речи и производится разбор на морфы в соответствии со списками морфов. При этом проверяется, чтобы после морфа

последней позиции во входном буквенном потоке следовал символ, который может служить разделителем слов. Полученный вектор (по позициям) номеров морфов через таблицу морфов преобразуется в вектор номеров морфем, после чего определяется вектор номеров масок и производятся проверки по маскам и номерам строк таблицы морфов, и затем на выход выдается конкатенация грамматем по всем позициям и номер лексемы.

Заметим, что приведенный алгоритм может быть легко использован для анализа сложных слов даже в том случае, когда априорно трудно указать место (места) членения композитной словоформы (*молокозавод*, *газоводопыленепроницаемый*, *тридцатидвухразрядный*). Такая необходимость может возникнуть при анализе текстов по химии или текстов на немецком или японском языке. В этом случае следует включить соединительную гласную (или иной соединительный морф, в том числе пустой) с грамматемой {композитность} в парадигмы лексем соответствующих частей речи, а в алгоритме анализа не требовать, чтобы после морфа с такой пометой во входном потоке следовал разделитель.

При синтезе выполняются преобразования, обратные по отношению к выполняемым при анализе. По инверсному файлу словаря определяется одна из основ лексемы с предъявленным номером. Желательно, чтобы это была частотно наиболее вероятная основа - например, первая из указанных лингвистом при формировании словаря. По ее словарной информации определяется техническая часть речи и, если возможно, набор масок. Производится разбор предъявленного набора характеристик на грамматемы; при этом просто выбираются все грамматемы, не противоречащие входному набору характеристик, и из них всеми способами составляются вектора по позициям. Производятся проверки их совместимости по таблицам, а также пригодность данного вектора по маскам, взятым из словаря. Для подошедших основ и векторов по таблице морфов определяются слева направо морфы (номер строки таблицы определяется уже построенными морфами). Конкатенация полученных морфов выдается на выход.

Оптимизации изложенных алгоритмов и внутренние структуры таблиц выбираются программистом.

НОРМАЛИЗАЦИЯ

При нормализации номерами лексем из алгоритма анализа можно исключить только заключительные операции выдачи характеристик. В этом случае можно полностью исключить из рассмотрения грамматемы и не хранить их в памяти, поскольку вся необходимая информация о синтактике морфем содержится в номерах масок, которые только и следует оставить в списках грамматем.

При нормализации словарными формами [5] после этого следует рассмотреть задачу синтеза «первой» формы. Однако выбор требуемого набора характеристик представляет некоторую проблему (ножницы, некого, щец, рад [10]). Один из путей решения этой проблемы состоит в следующем требовании, имеющем значение и для задачи синтеза независимо от задачи нормализации.

Фиксируем порядок перебора морфем (столбцов таблицы морфов) при синтезе, при котором вектора морфем синтезируются в лексикографическом порядке: при фиксированном номере морфемы позиции с меньшим номером перебираются слева направо столбцы таблицы морфов (т.е. сверху вниз грамматемы) позиций с большими номерами. Наложим на лингвистические таблицы дополнительное требование: при таком переборе синтезированные вари- анты словоформ должны появляться на выходе в «естественном» порядке (*красный*, *красного*, ..., *краснейший*, *краснейшего*, ...).

Опыт составления таблиц показывает, что при «естественном» перечислении парадигмы в каждой позиции указанное условие упорядоченности выдачи при синтезе в большинстве случаев выполняется автоматически.

Конкретно для целей нормализации требуется, чтобы в случае ущербности парадигмы первой по порядку синтезировалась именно та форма, которую можно считать словарной. Мы говорим о том, что форму «можно считать» словарной, поскольку для отождествления словоформ одной лексемы могла бы подойти любая форма. Итак, при нормализации словарными формами можно рассмотреть синтез всех форм лексемы (синтез с пустым условием на характеристики) и остановить его после нахождения первого варианта.

Технически такой алгоритм «синтеза сразу после анализа» или даже «синтеза в процессе цикла анализа» реализован автором с учетом того, что некоторая информация о синтезируемой лексеме (например, одна из ее основ) уже находится в памяти.

ИСПРАВЛЕНИЕ ОШИБОК

Алгоритм анализа построен так, что текстуальный разбор слова на морфы $m_0m_1...m_N$, так что морф m_i входит в список морфов i-й позиции (технической части речи, определяемой словарно по m_0), производится без учета какой-либо дополнительной информации, в частности, номеров строк таблиц морфов и синтактики морфов и морфем. В этом случае при анализе могут обнаруживаться и исправляться ошибки следующих трех типов:

1. Несовместимые грамматемы (**чита-ем-ый-ся*);
2. Недопустимая грамматема (**рад-ый*);
   Неверный выбор основы (**глуboч-ий*, **человек-и*);
3. Неверный выбор морфа (**красн-ий*).

В последних двух случаях по таблице морфов могут быть найдены все характеристики исходной словоформы и затем при необходимости синтезирована правильная. Необходимо отметить, что если таблицы не будут организованы специально с учетом задачи исправления ошибок, то такое исправление может ока- заться не всегда удачным.

При разбиении слов на технические части речи следует учитывать, что интерпретация и исправление ошибок действует только внутри одной части речи. Так, числительное **трух* (как *двух*) будет исправлено на *трех*, а существительное **методух* не будет исправлено на *методов*, поскольку морфа *-ух* нет в таблицах существительных.

Таким образом, программой будут «правильно» (т.е. так же, как человеком) поняты и исправлены словоформы вроде **пишат* вместо *пишут*, **красний*, **краснайшый*, **читающыйсь* и т.п. Кроме того, программа будет правильно понимать не существующие либо окказиональные синтетические словоформы в случаях, когда правильная словоформа должна быть аналитической (напр. **лучшайший* вместо *самый хороший*).

Заметим, что класс исправляемых ошибок существенно отличается от класса однобуквенных ошибок [11], часто рассматриваемого при обсуждении задачи исправления ошибок. В настоящее время готовится к публикации и описание алгоритма исправления однобуквенных ошибок в нашей модели [7].

ЗАКЛЮЧЕНИЕ

Описанная модель представляет собой по существу некоторый инструментарий для построения языково-независимой (в пределах определенного класса языков) программной оболочки системы морфологического анализа и синтеза. Структура представления лингвистической информации приближена к используемой во внутренних алгоритмах —

более, чем в системах, использующих общий механизм формальных грамматик [4], но менее, чем в системах без полного разделения программной и лингвистической информации. За счет этого повышено быстродействие морфоанализатора.

Исключив из рассмотрения чередования, мы стремились по возможности упростить алгоритм анализа, сделав его однонаправленным — «слева направо», причем поиск в словаре при этом можно рассматривать как просмотр дерева. Это позволило как применять методы, рассмотренные в [6], так и получить некото рые специальные эффекты, например, высокое быстродействие при массовом анализе сложных слов, раннее обнаружение ошибок при вводе с клавиатуры [7], эффективное исправление опечаток [7].

Рассмотрение членения словоформы на произвольное число позиций (суффиксов) позволяет формально описать как явления, относящиеся к словоизменению, так и некоторые явления, традиционно относимые к словообразованию [1, 2, 5].

Модель может быть использована в интеллектуальных системах понимания текста, поскольку при анализе строится полный набор характеристик и выдается уникальный и словарно закрепленный за лексемой номер.

Развитый в настоящей работе аппарат может быть также применен к следующим задачам: описание аномалий начертания некоторых букв, как, например, начертание русской буквы ё как е; описание расстановки ударений; приближенный анализ незнакомых слов.

Автор выражает благодарность проф. Г. Г. Белоногову, прочитавшему статью в рукописи и сделавшему ряд ценных замечаний. Научное руководство работой осуществлял проф. И. А. Большаков. На ранних стадиях работу консультировала по лингвистическим вопросам Е. Р. Добрушина. С. А. Крылов также прочитал статью в рукописи и оказал автору помощь в подборе терминологии. Различные варианты лингвистического обеспечения были составлены Н. В. Перцовым (ВНТИЦентр) и Н. М. Якубовой (ВИНИТИ).

Литература

1. Белоногов Г.Г., Зеленков Ю.Г. Алгоритм морфологического анализа русских слов // Вопросы информационной теории и практики. М.: ВИНИТИ, 1985, N 53.
2. Мальковский М.Г. Диалог с системой искусственного интеллекта. — М.: МГУ, 1985, 216 С.
3. Апресян Ю.Д. и др. Лингвистическое обеспечение системы автоматического перевода Этап-2. — М.: Наука, 1989.
4. Бидер И.Г., Большаков И.А., Еськова Н.А. Формальная модель русской морфологии. — ИРЯ АН СССР, Проблемная группа по экспериментальной и прикладной лингвистике. — вып. 112. — М. — 1978. — I — 60 С., II — 59 С.
5. Хорошилов А.А. Автоматическая нормализация слов в системах обработки научно-технической информации. — дисс. на степень кандидата технических наук. — М.: ВИНИТИ, 1987.
6. Гельбух А.Ф. Минимизация количества обращений к диску при словарном морфологическом анализе // Научно-техническая информация, серия 2. — М.: ВИНИТИ, 1991, N 6.
7. Гельбух А.Ф. Минимизация числа гипотез и обращений к дисковой памяти при словарном исправлении опечаток // Научно-техническая информация, серия 2. — М.: ВИНИТИ (в печати).
8. Лингвистический энциклопедический словарь. — М.: Советская энциклопедия, 1990, 685 стр.

9. Добрушина Е.Р., Савина Г.Б., Гельбух А.Ф. Система точного морфологического анализа и синтеза // Программное обеспечение новой информационной технологии. — Калинин: НПО ЦПС, 1989.
10. Зализняк А.А. Грамматический словарь русского языка. Словоизменение. — М.: Русский язык, 1980, 880 стр.
11. Большаков И.А. Минимизация перебора альтернатив при автоматизированном исправлении искаженных слов // Семиотика и информатика, вып. 31. — М.: ВИНИТИ, 1990.