

Министерство науки  
и технической политики  
Российской Федерации

Российская  
Академия наук

Всероссийский институт научной и технической информации

*На правах рукописи*

ГЕЛЬБУХ  
Александр Феликсович

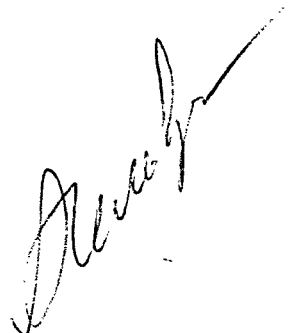
УДК [801.73:681.3] (043.3)

**ЭФФЕКТИВНО РЕАЛИЗУЕМАЯ НА ЭВМ  
МОДЕЛЬ МОРФОЛОГИИ  
ФЛЕКТИВНОГО ЕСТЕСТВЕННОГО  
ЯЗЫКА**

Специальность 05.13.17  
Теоретические основы информатики

**А В Т О Р Е Ф Е Р А Т**  
*диссертации на соискание ученой степени  
кандидата технических наук*

Москва 1994



Министерство науки  
и технической политики  
Российской Федерации

Российская  
Академия наук

Всероссийский институт научной и технической информации

*На правах рукописи*

ГЕЛЬБУХ

Александр Феликсович

УДК [801.73:681.3] (043.3)

**ЭФФЕКТИВНО РЕАЛИЗУЕМАЯ НА ЭВМ  
МОДЕЛЬ МОРФОЛОГИИ  
ФЛЕКТИВНОГО ЕСТЕСТВЕННОГО  
ЯЗЫКА**

Специальность 05.13.17

Теоретические основы информатики

**А В Т О Р Е Ф Е Р А Т**

диссертации на соискание ученой степени  
кандидата технических наук

Москва 1994

Работа выполнена во Всероссийском институте научной и технической информации.

Научный руководитель:  
доктор технических наук,  
профессор *Большаков Игорь Алексеевич*

Официальные оппоненты:  
доктор технических наук.  
профессор *Белонозов Герольд Георгиевич*,  
кандидат технических наук *Терещенко Сергей Сергеевич*

Ведущая организация: Российский Государственный Гуманитарный Университет.

Защита состоится *25.01* 1995 года в *12* часов на заседании Специализированного совета ДООЗ.02.01 во Всероссийском институте научной и технической информации по адресу: 125219, Москва, ул. Усиевича, д. 20-а.

С диссертацией можно ознакомиться в библиотеке Всероссийского института научной и технической информации.

Автореферат разослан *24.08* 1994 г.

Ученый секретарь  
Специализированного совета  
доктор технических наук

*Петрова*  
*Лидия Андреевна*

## ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы морфологического анализа и синтеза словоформ определяется тем, что блок морфологического анализа является необходимой частью большинства работающих с естественными языковыми текстами программ самого различного уровня и назначения; большинство таких систем нуждается также в блоке синтеза. Ввиду системного характера задачи и большого объема обрабатываемой информации к морфологическому блоку предъявляются жесткие требования по эффективности и быстродействию.

Задача заключается в том, чтобы разработать алгоритмы, методы и лингвистические модели, позволяющие автоматически осуществлять точный и полный морфологический анализ и синтез, а также решать ряд смежных задач, таких как нормализация слов, обучение пользователя грамматике, обнаружение и исправление грамматических ошибок и опечаток, интерпретация неправильных и неизвестных слов. К разрабатываемым алгоритмам предъявляется ряд более жестких, чем в известных алгоритмах, требований, в том числе - требование предельно высокой скорости работы и малого объема занимаемой оперативной памяти при работе на современных персональных ЭВМ, а также требование полного отделения программ от лингвистической информации в рамках модели, описывающей строение некоторого класса флективных языков.

Целью исследования в настоящей диссертации является разработка принципов, алгоритмов, программ и соответствующих лингвистических моделей, позволяющих создать эффективную по быстродействию и занимаемой оперативной памяти ЭВМ систему автоматического морфологического анализа, синтеза, нормализации слов, обнаружения и исправления ошибок, функционирующую на

современных персональных ЭВМ и допускающую встраивание в интегрированные пакеты обработки текстовой информации.

Предметом исследования является (1) изучение морфологического строения флективных языков, в частности, русского, в связи с задачей его формального описания в той мере, в какой это необходимо для построения программы автоматической морфологической обработки текста; (2) способы представления словаря и морфологической информации в связи с задачей ускорения доступа к хранящемуся на дисковом накопителе словарю; (3) алгоритмы морфологического анализа, синтеза, нормализации слов, обнаружения и исправления ошибок.

Научная новизна работы заключается в том, что автором впервые разработана структура словаря, позволяющая получить все гипотетические основы слова при предельно возможном быстродействии, то есть за одно элементарное обращение к дисковой памяти; разработана оригинальная языково-независимая (в некотором классе языков) модель морфологического строения флективного языка, основанная на разбиении словоформ на произвольное число равноправных в техническом отношении морфов; разработан метод исправления ошибок в тексте на флективном языке, превосходящий по быстродействию известные методы исправления ошибок данного класса; впервые предложен метод упорядочения процесса перебора альтернатив при исправлении опечаток, заключающийся в проверке гипотез в порядке возрастания времени, необходимого для каждой проверки.

Методы исследования. Исследование проводилось путем изучения закономерностей морфологического строения флективных языков, в первую очередь - русского; разработки конкретных морфологических таблиц и словаря для русского языка; изучения

методов представления словаря на устройстве дисковой памяти и алгоритмов доступа к нему; разработки алгоритмов морфологического анализа, синтеза, нормализации слов, обнаружения и исправления ошибок; практической реализации этих алгоритмов на ЭВМ; статистической обработки результатов экспериментов.

Практическая значимость работы заключается в том, что в результате проведенных исследований создана библиотека процедур, осуществляющих автоматический морфологический анализ, синтез, нормализацию слов, обнаружение и исправление ошибок на персональной ЭВМ. Данное программное средство позволяет существенно повысить эффективность реализованных на персональных ЭВМ диалоговых поисковых систем и систем подготовки документов, а также может служить инструментом дальнейшей лингвистической обработки текстов, включающей сбор различной статистики, поиск и выделение из текста фрагментов по различным условиям, синтаксический анализ и др.

#### Основные научные результаты:

- Разработана формальная модель морфологического строения флективного языка, позволяющая в некотором классе языков полностью отделить программы от лингвистических данных и допускающая эффективную реализацию на ее основе алгоритмов морфологического анализа, синтеза, нормализации слов, обнаружения и исправления ошибок.

- Разработана структура словаря, позволяющая достичь предельного быстродействия алгоритма поиска основ слов в словаре, а также алгоритмы поиска в таком словаре и алгоритм формирования словаря нужной структуры.

- Разработаны и практически реализованы на ЭВМ быстродействующие и экономичные по использованию оперативной памяти

алгоритмы морфологического анализа, синтеза, нормализации слов, обнаружения и исправления ошибок.

- Предложен новый метод упорядочения процесса перебора альтернатив при исправлении ошибок, заключающийся в генерировании в первую очередь тех гипотез, проверка которых требует минимального времени.

- При участии автора разработан и реализован на ЭВМ интерфейс стандартной библиотеки морфологических процедур.

- При участии автора создана система морфологических таблиц, описывающая строение русского языка в рамках разработанной автором модели.

- При участии автора создан машинный морфологический словарь русского языка (на основе известных источников), включающий около 130 тыс. лексем общеупотребительной и специальной лексики.

Реализация результатов работы. На основании разработанных алгоритмов, программ, лингвистического обеспечения и словаря создана стандартная библиотека процедур морфологического анализа, синтеза, нормализации слов, обнаружения и исправления ошибок. Имеется опыт интеграции данной библиотеки в реализованную на персональной ЭВМ информационно-поисковую систему DS-SIMPLE, практически эксплуатируемую с 1992 года. Получен также опыт реализации разработанного алгоритма исправления опечаток в системе грамматической проверки текста, основанной на иной морфологической модели.

Апробация работы. Основные научные результаты работы представлялись на конкурс работ молодых ученых ВНИИ Центра (1990; III место), докладывались автором на IV Всесоюзной школе-семинаре при Институте кибернетики АН УССР, конференции

"Программное обеспечение новой информационной технологии", III Международной конференции "Программное обеспечение ЭВМ", 1-й Ежегодной Всесоюзной конференции SUUG, конференции "Использование программных средств ЭВМ для автоматизации учредческой деятельности", Международном форуме "Тех-Екс'90 - обмен технологиями" (Болгария), на научно-технических семинарах сектора ПО баз знаний отдела автоматизации информационных процессов ВНИИЦентра и отдела ОТОИ ВИНТИ.

Публикации. По теме диссертации опубликовано 11 работ - три статьи, тезисы докладов, технические отчеты.

Структура и объем работы. Диссертационная работа состоит из введения, четырех глав и десяти приложений.

В первой главе обосновывается необходимость совершенствования методов морфологической обработки текстов на естественном языке. Дается обзор существующих подходов к решению проблемы, применяемых методов и разработанных к настоящему времени морфологических систем, анализируются их достоинства и недостатки. Формулируются задачи исследования.

Во второй главе излагается разработанный автором метод организации морфологического словаря, позволяющий при морфологическом анализе получать всю необходимую для анализа слова информацию при одном обращении к дисковой памяти. Приводятся алгоритмы формирования словаря и поиска информации в нем.

Третья глава является центральной в работе. В ней излагается разработанная автором языково-независимая в некотором классе естественных языков формальная лингвистическая модель морфологического строения естественного языка, а также структура разработанной автором морфологической системы, основанной на данной модели. Приводятся алгоритмы морфологического анали-



за, синтеза, нормализации слов, обнаружения, объяснения и исправления ошибок.

В четвертой главе отдельно рассматривается разработанный автором алгоритм исправления опечаток, примененный, в частности, в описываемой морфологической системе. Показывается, что по эффективности данный алгоритм превосходит применяемые в настоящее время. Приводится несколько вариантов алгоритма.

Общий объем работы - 220 страниц, основной текст - 200 страниц. Список литературы включает 171 название. В работе имеется 1 таблица и 9 рисунков.

#### КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В диссертации выбран следующий порядок изложения материала по главам.

Во введении обосновывается актуальность выбранной темы, определяется предмет, цель и методы исследования; раскрывается научная новизна и практическая значимость работы; формулируются основные научные результаты исследования; описывается структура диссертации и дается краткая характеристика ее основных разделов.

В первой главе раскрывается роль морфологической подсистемы в системах обработки естественных языковых текстов, информационного поиска и общения с пользователем на естественном языке. Обосновывается необходимость совершенствования существующих методов морфологического анализа, синтеза, грамматического контроля и обработки ошибочных слов.

Морфологической подсистемой называется комплекс программных средств, осуществляющих морфологические операции над

текстом на естественном, например русском, языке. Под морфологическими операциями понимается операции над текстом, связанные с явлением словоизменения (более точно, морфологического словоизменения), то есть образования различных форм слов. Например: *молоток* - *молотка* - *молотками*; *важный* - *важная* - *важен*; *читать* - *читали* - *читанное* - *читающимися* - *читая*. Образование форм слов свойственно подавляющему большинству крупных языков мира, однако наиболее сложный характер данное явление носит в языках так называемого флективного типа, к которому относится, в частности, русский язык.

Как в тексте, так и в информационных системах собственно слово, или лексема, и его морфологическая форма играют принципиально различную роль. Любые свойства понятия привязаны, как правило, к лексеме: так, с лексемой *МОЛОТОК* может быть связано свойство мужского рода, количество единиц на складе, цена, номера документов, содержащих определенную информацию, и т. д. С морфологической же формой, или набором грамматических характеристик, связана информация принципиально иного сорта: например, творительный падеж определяет роль непрямого дополнения в предложении. Поэтому при обработке текста принципиальное значение имеет задача разделения двух указанных сущностей и их раздельной обработки. Однако в языках флективного типа обе указанные сущности выражаются, как правило, одной буквенной цепочкой, называемой словоформой, ср.: *молотком*.

Морфологическим анализом называется сопоставление словоформе выражаемой ею лексемы, с одной стороны, и набора грамматических характеристик, с другой стороны. Например, *молотком* --> (1) *МОЛОТОК*, (2) творительный падеж единственного числа. Нормализацией слова называется сопоставление словоформе номера

лексемы, общего для всех словоформ данной лексемы, либо стандартизированной словарной формы, например: **МОЛОТКОМ** --> **МОЛОТОК**, **НОЖНИЦАМИ** --> **НОЖНИЦЫ**, **РАДЫ** --> **РАД**.

Морфологическим синтезом, называется обратное сопоставление, то есть нахождение словоформы по заданным лексеме и грамматическим характеристикам, например: **МОЛОТОК**, творительный падеж единственного числа --> **МОЛОТКОМ**.

Грамматической проверкой называется установление наличия в языке данной словоформы, включая проверку возможности выбора данных грамматических характеристик и правильности построения словоформы. Например, следующие словоформы ошибочны: **\*ножницей**, **\*хочут**, **\*слово**. Интерпретацией ошибочного слова называется решение задачи морфологического анализа даже для ошибочной словоформы: **\*ножницей** --> **НОЖНИЦЫ**, творительный падеж единственного числа. Объяснением ошибки называется установление природы ошибки или правила, запрещающего построение данной словоформы, например: "у лексемы **НОЖНИЦЫ** отсутствуют формы единственного числа". Исправлением ошибки называется установление правильного написания слова, например: **\*хочут** --> **хотят**, **\*слово** --> **слово**. Опечаткой называется грамматически немотивированная ошибка, например: **\*слово**, **\*прогрмма**.

Приближенным морфологическим анализом и синтезом называется решение соответствующих задач для словоформ, образованных от лексем, отсутствующих в словаре системы, например: **дистрибьюторами** --> **ДИСТРИБЬЮТОР (?)**, творительный падеж множественного числа (?); вопросительный знак здесь обозначает предположительность результата. Данная задача решается путем решения задачи приближенного определения морфологических признаков лексемы, то есть соответствующих ей правил образования морфологических форм.

Указанные задачи не являются тривиальными ввиду большого разнообразия словоизменительных классов во флексивном языке и сложности формальных правил преобразования буквенного состава слова при образовании его форм. Ср.: при анализе: *брусом* --> *брус*, *диском*, *буйком*, *партком*, *босиком*, *незнаком*; при синтезе: *СТОЛ* --> *столы*, *стулья*, *сны*, *сыновья*, *небеса*, *стога*, *круги*, *лотки*, *потоки*, *дети*. Построение списка всех словоформ языка в явном виде не приводит к эффективному решению задач морфологической обработки текста, хотя бы ввиду недопустимого замедления выборки информации из столь большого массива.

В функции универсальной морфологической системы входит автоматическое решение перечисленных выше задач. Поскольку морфологической обработке подвергается каждое слово текста в отдельности, основным техническим требованием к практически эксплуатируемой системе является требование высокого быстродействия. Кроме того, к универсальной системе предъявляется ряд дополнительных требований, таких, как требование последовательной обработки текста, учет переносов слов, обработка сложных слов и терминов, содержащих специальные символы, например: *тридцатидвухразрядный*, *OS/2 2.1*. В работе показано, что указанные задачи также не являются тривиальными.

Инструментальной называется система, допускающая встраивание в произвольную прикладную систему, нуждающуюся в морфологической обработке текста, и работу под ее управлением. Совместимость морфологической подсистемы с использующей ее прикладной программой накладывает на нее дополнительно ряд технических требований, в частности, требование минимизации объема используемой оперативной памяти ЭВМ; требование неизменности

номеров, назначаемых системой лексемам, при пополнении и изменении словаря; отделение лингвистической информации от программ и алгоритмов.

В области морфологической обработки текстов на естественном русском языке рядом исследователей достигнуты значительные результаты. Крупные практически эксплуатируемые системы созданы, в частности, научными коллективами под руководством Г. Г. Белоногова, И. А. Большакова, Ю. Д. Апресяна, М. Г. Мальковского, О. С. Кулагиной, С. А. Старостина. Значительные результаты достигнуты и зарубежными исследователями, например, К. Коскенниemi.

В работе дан подробный обзор существующих методов и систем морфологической обработки текста и обоснована необходимость их дальнейшего совершенствования и разработки высокоэффективной универсальной инструментальной морфологической системы, языковнезависимой в некотором классе языков, включающем русский язык. Показано, что применяемые в настоящее время системы не удовлетворяют перечисленным выше требованиям.

При разработке подобной системы необходим комплексный учет всей совокупности предъявляемых к ней требований. Так, требования высокого быстродействия и небольшого объема занимаемой оперативной памяти являются противоречивыми при использовании крупного морфологического словаря. Учет предъявляемых требований в их совокупности является определяющим фактором при разработке всех компонент системы, от лингвистической модели до структуры внутреннего представления морфологического словаря. Особое внимание, однако, уделяется наиболее важным и ресурсоемким задачам, в первую очередь - задачам морфологического анализа и исправления опечаток.

Во второй главе излагается способ организации внутреннего представления расположенного на дисковом накопителе морфологического словаря, обеспечивающий минимально возможное количество обращений к дисковому устройству на одно анализируемое слово, а именно - одно обращение на слово.

Выбор дискового накопителя для хранения крупного морфологического словаря обусловлен требованием минимизации используемой оперативной памяти. Дисковая память обладает определенными особенностями по сравнению с оперативной памятью ЭВМ, в частности, блочным принципом доступа. Последнее означает, что при одном обращении к дисковому устройству в оперативную память считывается определенный отрезок информационного массива - блок размером порядка одного килобайта, причем время поиска и считывания такого блока значительно превышает время обработки информации в оперативной памяти ЭВМ.

Рассматривается суффиксальное словоизменение, то есть тип словоизменения, при котором словоформы языка строятся из словарной основы и приписанного к ней справа и только справа набора суффиксов, например: *чит-а~~к~~ш~~и~~м~~и~~ся*, где *чит-* - основа, *-а-кш-им-ся* - тема + суффикс + окончание + частица. Чередования букв сводятся к включению в словарь различных основ одной лексемы. Словарь включает только основы лексем вместе с приписанной к ним грамматической и иной информацией. Предполагается, что словарь упорядочен по алфавиту.

Задача морфологического анализа, таким образом, сводится к отысканию в словаре всех начальных подцепочек предъявленной цепочки. Так, для слова *отними* необходимо проверить гипотетические основы *отними-*, *отним-*, *отни-*, *отн-*. Проблема заключается в том, что алфавитное место этих цепочек оказывается в

разных блоках словаря, что приводит к необходимости считывания в оперативную память с дискового накопителя нескольких блоков.

Предлагается продублировать в начало каждого блока словаря все словарные статьи с основами, являющимися начальными подцепочками первой основы данного блока. Например, в блок, начинающийся основой *паровоз-*, предлагается продублировать основы *паров-(ой)*, *пар-(а)*, *пар-(\*)*, *па-(й)*. Очевидно, что такое дублирование не приводит к существенному увеличению объема словаря.

Приводится алгоритм поиска основы слова в таком словаре, позволяющий найти все имеющиеся в словаре основы, являющиеся начальными подцепочками предъявленной буквенной цепочки, в одном блоке словаря.

Алгоритм не требует никакой лингвистической информации, в частности, информации о системе суффиксов и окончаний данного языка или о сочетаемости букв в слогах. Данное обстоятельство позволяет считать рассматриваемую задачу как частный случай задачи поиска информации в базе данных, а именно, как задачу нахождения в базе всех записей, ключи которых являются начальными подцепочками предъявленной буквенной цепочки.

Приводятся однопроходные алгоритмы формирования такого словаря из алфавитно упорядоченного набора словарных статей, а также обратного преобразования. Рассматривается также необходимый для решения задачи морфологического синтеза алгоритм нахождения в словаре лексемы по ее номеру, полученному ранее в процессе морфологического анализа.

Показывается, что для решения одной важной частной задачи, относящейся к проблеме поиска ближайшей цепочки и существенно используемой при решении задачи исправления орфографи-

ческих ошибок, также достаточно одного обращения к дисковой памяти.

В третьей главе, центральной в работе, излагается модель морфологии флективного естественного языка, являющейся основой высокоэффективной программной реализации алгоритмов морфологического анализа, синтеза и других приведенных выше морфологических операций. Модель является языково-независимой в некотором классе языков, включающем русский, что обеспечивает отделение лингвистической информации от программ и алгоритмов.

Рассматриваются требования, предъявляемые к лингвистической модели, используемой в универсальной морфологической системе, и обосновываются принятые при построении модели конструктивные решения.

Отправной точкой при построении модели служит приведенный во второй главе алгоритм поиска в морфологическом словаре всех гипотетических основ словоформы без привлечения лингвистической информации. Соответственно, модель описывает суффиксальное словоизменение, причем любая информация о связи различных частей слова привязывается к более левой части и влияет на выбор более правых частей. Другими словами, все правила в модели записываются в виде "после того-то пишется то-то", но не "перед тем-то пишется то-то".

Словоформы языка в модели образуются путем соединения буквенных цепочек - флексий, или формальных морфов, с каждой из которых связан набор морфологических характеристик. Морфы сочетаются в словоформах языка в соответствии с определенными закономерностями. Первый из морфов словоформы, основа, определяет лексему и является специфическим для данной лексемы,



остальные являются стандартными для данного языка. Например, читающийся:

чита- --> лексема=ЧИТАТЬ, часть речи=глагол,  
-ющ- --> залог=действительный, время=настоящее,  
-ий- --> падеж=имен., род=мужск., число=единств.,  
-ся --> возвратный.

Соответственно, лингвистическая информация задается для основ - в хранящемся на диске словаре, для остальных морфов - в расположенных в оперативной памяти лингвистических таблицах. При этом задается информация трех типов.

Во-первых, задается соответствие между морфами и морфологическими характеристиками. Например, падежи некоторого класса существительных выражаются окончаниями -\*, -а, -у, -\*, -ом, -е. Или: чита- --> часть речи = глагол.

Во-вторых, задается соответствие между буквенными вариантами начертания морфов, задаваемое в виде правил вида "после того-то пишется то-то". Например, после основы мужского рода пал- пишутся окончания -а, -ы, -е, -у, -ой (-ою), -е. Или: после -ющ- пишется -ий, -его, ..., -ем, но после -ем- пишется -ый, -ого, ..., -ом: чи-та-ющ-ий, чита-ем-ый.

В-третьих, соответствие между грамматическими характеристиками, выражаемыми различными морфами, задаваемое в виде правил "после того-то можно/нельзя/нужно употреблять то-то". Например, после суффикса, выражающего страдательный залог причастия, нельзя употреблять частицу, выражающую возвратность: \*чита-ем-ый-ся. Аналогичные правила задаются для основ. Например, после основы смея- необходимо употреблять возвратную частицу: смея-ть-ся. Однако при этом выбор конкретного буквен-

ного варианта частицы определяется правилами первой группы: *смея-ла-сь*.

Вся лингвистическая информация задается в виде таблиц, а указанные выше правила - в виде ссылок на строки таблиц, приписываемых основам, морфам и столбцам таблиц. В модели имеются предопределенные правила разрешения конфликтов при определении элементов словоформы. Например, основа *важн-* задает окончание *-ый*, а суффикс *-ейш-* задает окончание *-ий*. Выбирается наиболее правый управляющий элемент: *важн-ейш-ий*.

Кроме возможности или невозможности построения определенной формы слова, учитывается также затрудненность, факультативность и другие особенности употребления словоформы. Лингвистическая информация задается в форме, симметричной относительно процессов анализа и синтеза словоформ.

Приводятся алгоритмы морфологического анализа, синтеза, нормализации, интерпретации и исправления некоторых типов грамматически мотивированных орфографических ошибок. Большое внимание уделяется проблеме правильной обработки переноса слов при анализе текста. Учитывается заглавность букв. Приводятся технические приемы сокращения расхода оперативной памяти ЭВМ и времени поиска информации в словаре. Приводятся алгоритмы работы одновременно с несколькими словарями, а также в режиме совместного анализа и синтеза словоформ, например, при нормализации слов или при автоматическом переводе.

В четвертой главе рассматривается отдельно алгоритм исправления опечаток, то есть грамматически немотивированных орфографических ошибок, базирующийся на предложенных в предыдущих главах алгоритмах поиска информации в словаре и морфологического анализа. В связи с большим числом испытываемых вариан-

тов процедура исправления опечаток предъявляет наиболее жесткие требования к вычислительной эффективности алгоритма морфологического анализа, а также является наиболее медленной частью морфологической системы. Быстродействие существующих методов исправления опечаток не является удовлетворительным.

Излагаемый в работе алгоритм минимизирует число обращений к дисковой памяти в процессе исправления опечаток. Поскольку именно считывание с диска одного блока словаря является наиболее медленной операцией в процессе морфологического анализа, такая минимизация необходима для сокращения времени работы алгоритма в целом.

Минимизируются совместно следующие четыре показателя:

(1) время получения первого варианта исправления; (2) время получения правильного варианта; (3) время получения последнего из найденных вариантов; (4) общее время поиска, то есть время установления того факта, что все варианты найдены исчерпывающе. Рассматриваются как средние значения, так и медианы соответствующих распределений. Особенностью алгоритма является получение, как правило, всех вариантов исправления на самой ранней стадии работы алгоритма, составляющей около 2% общего времени работы.

Предложенные алгоритмы ориентированы в основном на исправление так называемых однобуквенных опечаток - вставок, замен, пропусков одной буквы и перестановок соседних букв. Приводятся также модификации алгоритма, осуществляющие исправление неоднобуквенных и неединичных ошибок.

Алгоритм основан на принципе обхода "вглубь" дерева, представляющего упорядоченный по алфавиту словарь. Однако в целях минимизации первых трех из указанных выше показателей

применяется специальный прием, а именно, перебор испытуемых позиции букв в слове осуществляется не от начала слова к концу, как в обычном алгоритме обхода дерева, а от конца слова к началу. Указанный прием позволяет более чем на порядок сократить время получения вариантов исправления, практически без увеличения общего времени работы алгоритма.

Приводимый алгоритм не использует никаких дополнительных лингвистических или иных данных, кроме словаря и одной из таблиц, используемых процедурой морфологического анализа. Реализующие алгоритм программы имеют небольшой размер и практически не используют дополнительной по сравнению с процедурой морфологического анализа оперативной памяти.

Приводится также модификация алгоритма, опирающаяся на небольшой автоматически создаваемый по словарю массив данных и использующая в дополнение к рассмотренным выше методам некоторый вариант так называемого метода  $n$ -граммного контроля, что позволяет достичь существенного сокращения общего времени работы алгоритма. Приводится также модификация алгоритма, осуществляющая почти полный (но не полный) поиск вариантов исправления; при этом общее время работы алгоритма сокращается в 60 раз. Обсуждается проблема ранжирования вариантов исправления по вероятности оказаться правильным.

Результаты экспериментов показывают, что предложенный метод исправления опечаток выбранного типа по эффективности превосходит известные.

## ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

В итоге работы над темой диссертации решена поставленная в начале исследования задача разработки принципов, моделей, методов и алгоритмов, составляющих в совокупности универсальную, языковнезависимую в некотором классе языков, быстродействующую, экономичную по расходу оперативной памяти ЭВМ морфологическую систему, удовлетворяющую ряду дополнительных требований.

В процессе исследования получены следующие научные результаты:

1. Разработана модель морфологического строения флективного естественного языка, эффективно реализуемая на современных ЭВМ. Под эффективностью реализации понимается небольшая потребность в оперативной памяти в сочетании с высоким быстродействием. Модель является языково-независимой в некотором классе языков, включающем русский, и отвечает определенной совокупности предъявленных требований.

2. На основании предложенной модели в комплексе и с единых позиций разработаны и программно реализованы высокоэффективные алгоритмы решения задач точного и полного морфологического анализа и синтеза, а также ряда смежных задач, таких как нормализация слов, обнаружение, интерпретация и исправление ошибок и опечаток, приближенный анализ новых слов. Указанные алгоритмы и программы в совокупности составляют универсальную инструментальную языковнезависимую в некотором классе языков морфологическую систему.

3. При участии автора созданы система морфологических таблиц для русского языка в рамках предложенной модели, а также машинный морфологический словарь русского языка (на базе известных источников), включающий около 130 тыс. лексем общепотребительной и специальной лексики.

4. Применительно к задаче морфологического анализа и синтеза предложена структура базы данных, ориентированной на высокоэффективное решение следующей задачи: найти все записи, ключи которых являются начальными подцепочками предъявленной не ограниченной справа буквенной цепочки. Для нахождения всего искомого множества записей достаточно одного обращения к дисковой памяти, что является пределом возможного быстрого действия для крупного словаря при ограничении занимаемого объема оперативной памяти. Разработаны однопроводные алгоритмы формирования и распаковки такой базы данных.

5. Предложены алгоритмы поиска вариантов исправления опечаток в слове, взятом вне контекста, с помощью перебора, управляемого морфологическим словарем. При этом минимизируются совместно следующие показатели: время нахождения первого, правильного, последнего вариантов исправления и время завершения процесса поиска. Особенностью алгоритмов является получение, как правило, всех вариантов исправления на самой ранней стадии работы алгоритма, составляющей около 2% общего времени работы. Экспериментально показано, что по эффективности разработанные алгоритмы превосходят известные.

6. Варианты высокоэффективного алгоритма исправления опечаток предложены, в частности, для решения задачи исправления одиночных однобуквенных опечаток и неоднобуквенных опечаток некоторого (произвольного) класса, а также задачи исправления различных типов неодинокных опечаток.

7. Созданы программные средства, реализующие перечисленные выше алгоритмы. Программы составлены в основном на языке С++ и содержат более 20 000 строк исходного текста.

Основные результаты диссертации отражены в следующих публикациях:

1. Гельбук А. Ф. Минимизация количества обращений к диску при словарном морфологическом анализе // НТИ, сер. 2, 1991, N 6.

2. Гельбук А. Ф. Минимизация числа гипотез и обращений к дисковой памяти при словарном исправлении опечаток // НТИ, сер. 2, 1993, N 5, с. 23-30.

3. Гельбук А. Ф. Модель морфологии флективного естественного языка // Материалы III Международной конференции "Программное обеспечение ЭВМ", Тверь, НПО Центрпрограммсистем, 1990, с. 27-31.

4. Гельбук А. Ф. Морфологический анализ/синтез и проверка русских текстов // Тезисы 1-й Ежегодной Всесоюзной конференции SUUG, Москва, 1990.

5. Гельбук А. Ф. Простая оболочка системы точного морфологического анализа и синтеза текстов на естественном языке // Использование программных средств ПЭВМ для автоматизации учебно-научной деятельности, АН СССР и НПО ЦПС, Калинин, 1990.

6. Гельбук А. Ф. Пустая языково-независимая оболочка системы точного морфологического анализа и синтеза текстов на естественном языке // Международный форум "Тех-Екс'90 - обмен технологиями", Болгария, Пловдив, 1990.

7. Гельбук А. Ф. Эффективно реализуемая модель морфологии флективного языка // НТИ, сер. 2, 1992, N 1.

8. Добрушина Е. Р., Савина Г. Б., Гельбук А. Ф. Разработка программных и лингвистических средств для создания баз знаний

о научных исследованиях и разработках. Отчет ВНИИЦ N 02900 056145, Москва, 1989, 35 с.

9. Добрушина Е. Р., Савина Г. Б., Гельбук А. Ф. Система точного морфологического анализа и синтеза // Программное обеспечение новой информационной технологии, АН СССР и НПО ЦИО, Калинин, 1989.

10. Загацкий Б. А., Перцов Н. В., Гельбук А. Ф. Лингвистический процессор базы знаний о научных исследованиях и разработках. Отчет ВНИИЦ N 02900 003868, Москва, 1990, 82 с.

11. Загацкий Б. А., Перцов Н. В., Гельбук А. Ф. Система синтаксического анализа фраз русского языка // Международный форум "Тех-Екс'90 - обмен технологиями", Болгария, Пловдив, 1990.