



**Современный
Гуманитарный
Университет**

Дистанционное образование

Рабочий учебник

Фамилия, имя, отчество _____

Факультет _____

Номер контракта _____

ЛИНГВИСТИЧЕСКИЕ ОСНОВЫ ИНФОРМАТИКИ

ЮНИТА 1

ЯЗЫК КАК ЗНАКОВАЯ СИСТЕМА

МОСКВА 2000

Разработано В.И. Киселевым

Рекомендовано Министерством
общего и профессионального
образования Российской Федерации
в качестве учебного пособия для
студентов высших учебных заведений

КУРС: ЛИНГВИСТИЧЕСКИЕ ОСНОВЫ ИНФОРМАТИКИ

Юнита 1. Язык как знаковая система.

Юнита 2. Основы теории формальных языков и языки обработки
данных информационных систем.

ЮНИТА 1

Для студентов Современного Гуманитарного Университета

Юнита соответствует профессиональной образовательной программе №

ОГЛАВЛЕНИЕ

ДИДАКТИЧЕСКИЙ ПЛАН	4
ЛИТЕРАТУРА	5
ТЕМАТИЧЕСКИЙ ОБЗОР	6
1. Информационные системы и лингвистика	6
1.1. Информационное общество	6
1.2. Информация и информационные технологии	7
1.3. Развитие информационных систем	8
1.4. Роль и место лингвистики в информатике	10
1.5. Язык как знаковая система	12
1.5.1. Основные функции языка	12
1.5.2. Информация и знак, знаковые системы и представление языка	12
2. Структура естественных языков	14
2.1. Слово в естественном языке	14
2.2. Словосочетание в естественном языке	32
2.3. Предложение	36
3. Семантико-синтаксическая структура и анализ текстов	41
3.1. Семантико-синтаксическая структура текстов	41
3.2. Введение в семантико-синтаксический анализ текстов ..	46
3.3. Морфологический анализ и синтез слов	49
3.3.1. Общее описание морфологического анализа и синтеза	49
3.3.2. Флексивный анализ и синтез	51
3.3.3. Морфологический анализ и синтез слов с изменяемой основой	59
3.3.4. Алгоритмы морфологического анализа и синтеза ..	64
3.3.5. Сравнение различных методов анализа и синтеза ..	67
3.3.6. Многоступенчатый морфологический анализ и синтез	69
3.4. Анализ и синтез именных словосочетаний	70
3.4.1. Синтаксический анализ именных словосочетаний ..	70
3.4.2. Кодирование и декодирование наименований понятий	73
3.5. Синтаксический анализ текстов	75
ЗАДАНИЯ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ	77
ГЛОССАРИЙ*	

* Глоссарий расположен в середине учебного пособия и предназначен для самостоятельного заучивания новых понятий.

ДИДАКТИЧЕСКИЙ ПЛАН

Информационные системы и лингвистика. Информационное общество. Информация и информационные технологии. Развитие информационных систем. Роль и место лингвистики в информатике. Язык как знаковая система. Основные функции языка. Информация и знак. Знаковые системы и представление языка. Структура естественных языков. Слово в естественном языке. Словосочетание. Предложение. Семантико-синтаксическая структура и анализ текстов. Семантико-синтаксическая структура текстов. Введение в семантико-синтаксический анализ текстов. Морфологический анализ и синтез слов. Общее описание морфологического анализа и синтеза. Флективный анализ и синтез. Морфологический анализ и синтез слов с изменяемой основой. Алгоритмы морфологического анализа и синтеза. Сравнение различных методов анализа и синтеза. Многоступенчатый морфологический анализ и синтез. Анализ и синтез именных словосочетаний. Синтаксический анализ именных словосочетаний. Кодирование и декодирование наименований понятий. Синтаксический анализ текстов.

ЛИТЕРАТУРА

Базовая

1. Криницкий Н.А., Миронов Г.А., Фролов Г.Д. Автоматизированные информационные системы. М., 1982.

Дополнительная

2. Фостер Дж. Автоматический синтаксический анализ. М., 1975.
3. Попов Э.В. Общение с ЭВМ на естественном языке. М., 1982.
4. Ершов А.П. Введение в теоретическое программирование. Беседы о методе. М., 1977.

Примечание. Тематический обзор составлен на основе всех указанных источников.

Современный Гуманитарный Университет

1. ИНФОРМАЦИОННЫЕ СИСТЕМЫ И ЛИНГВИСТИКА

1.1. Информационное общество

В настоящее время информатика – одна из важнейших и перспективнейших «точек роста» мировой цивилизации, а широкое внедрение информационных и коммуникационных технологий во все сферы человеческой деятельности создает реальную основу называть общество наступающего XXI в. «информационным обществом».

Происходящий на наших глазах глобальный процесс формирования новой автоматизированной информационной среды общества создает беспрецедентные возможности для развития человека и эффективного решения многих профессиональных, экономических, социальных и бытовых проблем. Использовать эти возможности смогут лишь те члены общества, которые будут обладать необходимыми знаниями и умениями ориентироваться в новом информационном пространстве.

Представление о новом «революционном» прорыве человеческого общества в области технологий, как о переходе к «постиндустриальному, информационному обществу», появилось в 80-х гг. в результате анализа статистических данных о значительных изменениях в структуре занятого населения.

В условиях радикального усложнения жизни общества, его технической и социальной инфраструктуры, происходит изменение отношения людей к информации, которая становится таким же стратегическим ресурсом общества, как продукты питания, материальные или энергетические ресурсы.

Информационное (постиндустриальное) общество – общество, основным фактором развития которого являются автоматизированные информационные технологии.

«Информационная революция», решая одни проблемы, порождает новые.

Одной из таких проблем является «информационный взрыв», избыток доступных многим современным людям данных, которых больше, чем в состоянии переварить человеческое сознание, служит причиной снижения качества мышления среди членов современного общества. Информационная перегрузка – это реальность. Ежегодно публикуются десятки тысяч научных статей. Специалисты жалуются, что они не в состоянии оценить всего, что относится к их предметной области. Огромные объемы информации, особенно собранные в виде статистических данных, являются полем для ошибочной и (или) преднамеренно ложной интерпретации.

* Жирным шрифтом выделены новые понятия, которые необходимо усвоить. Знание этих понятий будет проверяться при тестировании.

Различие в уровне информационного обеспечения сегодня становится одной из существенных причин в дисбалансе экономического развития передовых и слабо развитых стран, порождает нестабильность в отношениях между странами. Знания – вся совокупность полезной информации и процедур, которые можно к ней применить, чтобы произвести новую информацию, – всегда дают власть тем, кто ими владеет и умеет ими пользоваться. Важное значение приобретает информационная культура – совокупность методов, приемов и навыков по сбору, хранению, обработке и созданию информации.

1.2. Информация и информационные технологии

Понятия «информация», «знание», «информационная система» следует в значительной степени считать интуитивными. До настоящего времени формальной теории автоматизированных информационных систем, подобной, допустим, теории кодирования и передачи информации или аналитической теории алгоритмов, не существует. В то же время существует инженерная дисциплина, которая базируется на хорошо формализованных теоретических построениях и имеет огромное значение в жизни современного общества. Эта дисциплина в различных странах имеет различные названия, но в содержательном плане ее предметная область является общепризнанной в мировом научном сообществе. В Европе эту дисциплину называют информатика, а в США и других англоязычных странах – компьютерная наука. Основными понятиями этой дисциплины являются следующие:

информация – любой вид сведений о предметах, фактах, понятиях предметной области, или сведения, не известные до их получения, являющиеся объектом хранения, передачи и обработки;

данные – представление информации в формализованном виде, удобном для пересылки, сбора, хранения и обработки;

сигнал – носитель данных (информации), может представлять собой физические сигналы или математические модели;

информационная технология (ИТ) – система научных и инженерных знаний, а также методов и средств, которая используется для создания, сбора, передачи, хранения и обработки информации в предметной области;

автоматизированная информационная технология (АИТ) – информационная технология, в которой для передачи, сбора, хранения и обработки данных используются методы и средства вычислительной техники и систем связи.

Информационные технологии имеют ряд замечательных свойств, присущих только им. Ниже перечислены основные из этих свойств, которые позволяют значительно увеличить эффективность всех процессов социально-экономического развития:

– современный научно-технический уровень ИТ таков, что им могут быть “перепоручены” практически все “нетворческие” процессы

обработки информации, которые используются в человеческой деятельности;

- основная масса навыков и приемов, которые используются не в творческих областях деятельности, могут быть автоматизированы;
- современные средства связи, охватывающие весь земной шар, позволяют обеспечить доступ к различным видам ИТ в любой его точке;
- на протяжении последних лет наблюдается неуклонное уменьшение абсолютной величины отношения стоимости к производительности у технических компонент автоматизированных информационных систем, что повышает экономическую эффективность использования ИТ.

Распространение телекоммуникационных технологий на базе средств вычислительной техники и средств связи вовлекает в сферу информационного общества все новые слои населения и новые виды человеческой деятельности – от бытовых до общественно-политических.

Таким образом, информационные технологии являются средством, которое позволяет достигать определенные цели, поставленные в различных областях человеческой деятельности, а также повышают эффективность решения задач, направленных на достижение этих целей.

Информационные технологии принято разделять на базовые и прикладные.

Под базовыми информационными технологиями понимается система научных и инженерных знаний, а также методов и средств, которая используется для создания, сбора, хранения и обработки информации безотносительно к предметной области, в которой создается и используется данная информация. В состав этих технологий входят:

- операционные системы;
- языки программирования и технологии их использования (компиляторы, библиотеки, CASE-технологии и т.д.);
- системы управления базами данных;
- технологии работы в телекоммуникационных сетях;
- экспертные и другие интеллектуальные системы.

С помощью базовых ИТ, средств вычислительной техники и связи разрабатываются и поддерживаются прикладные ИТ, обеспечивающие автоматизированную поддержку конкретных областей человеческой деятельности.

1.3. Развитие информационных систем

Процессы обработки информации всегда являлись основой человеческой деятельности, и объединение таких процессов с информационными ресурсами со временем стали называть **информационными системами (ИС)**. **ИС** – это комплекс, состоящий из информационной базы (хранилища информации) и процедур, позволяющих накапливать, хранить, корректировать, осуществлять поиск, обработку и выдачу

информации. С появлением вычислительной техники ИС пережили качественный процесс развития, превратившись в **автоматизированные информационные системы (АИС)**, т.е. – информационные системы, имеющие следующие основные компоненты:

физическая компонента – материальная основа носителя информационной системы;

информационная компонента – организованная определенным образом система записей данных (информационная база), характеризующаяся определенным языком, на котором выполнены образующие ее записи;

функциональная компонента – система процедур управления, обновления, поиска и завершающей обработки данных.

Современные АИС представляют собой чрезвычайно сложные человеко-машинные комплексы, интегрированные (неразрывно связанные) в национальную и мировую информационные среды. Эффективность АИС во многом определяется их качеством и доверием к ним пользователей. Качество изделий, процессов проектирования, производства и услуг является одной из узловых проблем, определяющей уровень жизни человека и состояние народного хозяйства, что полностью относится и к области информационных технологий. В АИС входят следующие основные составляющие:

- аппаратные средства вычислительной техники;
- аппаратные средства телекоммуникации (связи);
- программные средства реализации функций АИС;
- информационные базы данных (БД);
- документация, регламентирующая функции и применение АИС.

Аппаратное обеспечение АИС имеет достаточно универсальный характер и относительно слабо зависит от функционального назначения конкретной информационной технологии. Остальные компоненты АИС составляют их интеллектуальную часть, определяющую назначение, функции и качество решения задач в конкретной области человеческой деятельности. Эти компоненты могут отличаться принципиальной новизной, большим разнообразием характеристик, которые трудно формализуются и требуют глубокого исследования методов проверки их значений.

Любая реальная АИС действует в окружающей ее информационной среде, которую часто называют инфраструктурой.

Под инфраструктурой в экономике понимаются структуры, которые обеспечивают функционирование производственных систем, но непосредственно в технологических процессах производства продукции не участвуют. В их число входят: дороги, линии электропередачи, системы снабжения ресурсами и т.д.

Под инфраструктурой автоматизированных информационных систем обычно понимают телекоммуникационные сети и связываемые ими объекты: серверы, автоматизированные рабочие места, каталоги сетевых информационных ресурсов и т.п. Информационными

ресурсами являются информационные базы (банки и базы данных) различного назначения и другие информационные структуры.

В настоящее время особенно быстро развиваются телекоммуникационные компоненты инфраструктуры АИС. В области технологий передачи данных открываются новые горизонты в использовании сетей. При создании сетей передачи данных нового поколения возникает необходимость решения сразу комплекса задач, лежащих как в области телекоммуникаций, так и в области информационных технологий.

Быстрое развитие и использование информационных технологий не только открывает новые возможности, но и создает новые проблемы перед мировым сообществом, которые безусловно влияют на экономическую, социальную, культурную и образовательную деятельность. Такими проблемами являются:

- психо-биологические, оказывающие отрицательное психологическое и физическое воздействие на пользователей;
- культурные, угрожающие национальной культурной самобытности пользователей;
- социально-экономические, увеличивающие неравенство возможностей получения доступа к качественным ИТ;
- политические, способствующие разрушению гражданского общества в национальных государствах;
- бесконтрольное и несанкционированное использование чужой интеллектуальной собственности;
- технологические угрозы нанесения ущерба или разрушения самим АИС.

1.4. Роль и место лингвистики в информатике

Огромные средства затрачиваются во всем мире на разработку многочисленных, конкретных прикладных систем и совершенно недостаточное внимание уделяется теоретическим вопросам. Необходимость концептуального, системного осмысления положения дел в некоторой предметной области возникает в силу того, что на определенном этапе развития этой области накапливается большое количество знаний, фактов, задач и интересов, которые слабо увязаны между собой. Такое положение дел периодически возникает во всех развивающихся областях человеческих знаний, в быстро развивающихся областях процессы могут принимать кризисный характер.

В частности, при разработке сложных перспективных автоматизированных информационных систем процесс создания включает большое разнообразие видов деятельности и требует тесного взаимодействия между представителями научно-технических профессий и лицами, принимающими политические и экономические решения. Возникает необходимость сведения воедино огромных объемов разнообразной информации, согласования большого числа различных и зачастую противоречивых целей и интересов.

Этими обстоятельствами и вызвана необходимость закладывать в основу инженерной дисциплины «Автоматизированные информационные системы» хорошо проработанные разделы фундаментальных и прикладных научных дисциплин. Так как компоненты АИС принадлежат к различным областям знаний, то и теоретические основы АИС включают в себя положения из различных научных дисциплин.

При анализе и проектировании дискретных электронных схем для вычислительной техники используют математический аппарат теории множеств, двоичной логики (булева алгебра) и теории кодирования. При разработке цифровых систем связи используются методы математической статистики.

В разработках программного обеспечения и языков программирования используются методы автоматически-лингвистических моделей, аналитическая теория алгоритмов, модели исчисления предикатов, оптимизационные методы.

При разработке баз данных и информационных языков поиска и запроса данных в информационных базах используются методы реляционной алгебры, модели исчисления предикатов и математической лингвистики.

При анализе эффективности, работоспособности и надежности компонентов АИС и АИС в целом используются методы теории массового обслуживания и математическая статистика.

Наконец, в информационно-поисковых АИС используются методы теории информации, основными понятиями которых являются документ – материальный объект с информацией, закрепленной созданным человеком способом для ее передачи во времени и пространстве, и классификатор – официальный документ, представляющий систематизированный свод наименований и кодов кодификационных группировок и (или) объектов классификации.

Лингвистическое обеспечение АИС – совокупность языковых средств для формализации естественного языка, построения и сочетания информационных символов при общении пользователей с АИС – представляет собой, как видно из вышесказанного, один из основных видов базового научного обеспечения. Напомним, что **лингвистика** – наука о языке, общих законах строения и функционирования языка, а **информационный символ** – символ сообщения (записи), который является частью его содержания, в отличие от служебных (управляющих, разделителей) символов.

Лингвистические методы используются для **представления знаний в информационных системах** – формализации процедур, используемых биологическими объектами при решении интеллектуальных задач и при создании **интерфейсов пользователя** – средств, которые обеспечивают взаимодействие пользователя с АИС. В частности, в лингвистических моделях используются в моделировании знаний **фреймы** – схемы представления знаний, описывающие понятие или объект. Фрейм состоит из ссылки на суперпонятие (родовое понятие) и описания свойств, отличающих данный объект от суперпонятия.

1.5. Язык как знаковая система

1.5.1. Основные функции языка

Язык – форма существования знания в виде системы знаков плюс правила функционирования этих знаков, служащая средством человеческого общения, мышления и выражения. Таким образом, язык имеет две **основные функции** – он является орудием человеческого мышления и средством общения людей друг с другом.

Общение представляет собой двусторонний процесс передачи информации с определенными целями и по определенным правилам, выраженной на языке, понятном участникам общения, при этом под **понятностью** понимаются синтаксическая, семантическая и прагматическая однозначность информации, требующая общности языка и знаний участников о предметной области общения.

Кроме этого, язык обладает и **моделирующей функцией**, то есть обеспечивает представление некоторых характеристик физической или абстрактной системы средствами конкретного языка.

Основой моделирования является **экспликация** – строгая (математическая) формулировка содержательного или интуитивного понятия в предметной области моделирования.

В автоматизированных информационных системах средством моделирования служит **ограниченный естественный язык (ОЕЯ)** – искусственный язык, разработанный для общения человека с ЭВМ (диалект естественного языка).

Появление развитых диалоговых языков общения с ЭВМ еще более усилило интерес исследователей к функциям естественных языков. Для человека идеальным было бы общение с ЭВМ на обычном языке. Кроме того, естественный язык – это единственно известная на сегодня моделирующая система, средствами которой можно описать все многообразие окружающего мира. Основными задачами в области исследований языка как средства для описания действительности и средства общения между человеком и ЭВМ являются следующие:

- автоматизированный анализ текстов на естественном языке;
- переход от языковых представлений к языку описания знаний;
- понимание вопросов и формирование ответов в автоматизированных системах;
- извлечение знаний из текстов на естественном языке.

1.5.2. Информация и знак, знаковые системы и представление языка

В языковых исследованиях важное значение придается понятию **знак** – способу обозначения определенного понятия, предмета, свойства, используемому для приобретения, хранения, обработки и передачи информации. В частности, Ф. де Сосюр (выдающийся

лингвист) рассматривал язык как **знаковую систему**, – знаки ее не функционируют независимо друг от друга, а образуют систему, правила которой определяют закономерности их построения, осмысления и употребления (грамматика, правила смысла).

Этот подход затем развился в отдельную научную дисциплину, семиотику – науку о том, с помощью каких средств в человеческом обществе происходит общение (передача информации), в том числе:

- как устроены эти средства сами по себе;
- как они применяются;
- каким изменениям они подвержены и т.д.

Во всяком общении можно выделить смысл и средства его передачи. Если расчленить этот смысл на элементы и найти средства, которыми выражается каждый из них, – это и будут знаки, соединения определенного смысла и определенного способа его выражения (означаемого и означающего). Носителями смысла могут быть действия, понятия, предметы (идеальные и материальные). Например, выделяют следующие типы знаков:

знак-копия – воспроизведения, более или менее сходные с обозначаемым;

знак-признак – связан с обозначаемым предметом как действие со своими причинами (симптомы, приметы и т.п.);

знак-символ – наглядные образы, используемые для выражения некоторого, часто весьма значительного и отвлеченного, содержания (маска – символ театра).

В исследованиях знаковых систем применяется ряд специфических понятий; так, под **синтагмой** понимают сложный языковой знак (обычно двучленный), составленный из слов или морфем, соединенных определенным типом связи. Методы изучения синтагм составляют предмет **синтагматики**.

Синтактика занимается изучением структурных аспектов сочетаний знаков данной системы, правила их образования и преобразования безотносительно к их значениям и функциям, в то время как **прагматика** – изучением отношения, воспринимающего знаковую систему (интерпретатор или адресат), к самой знаковой системе.

Предмет, обозначаемый знаком, называется **денотат (референт)**, а **концепт** – это информация, которую знак несет о возможных денотатах, об их положении в системе реалий, об их месте в универсуме.

Экстенционал знака – определяет класс всех допустимых для этого знака денотатов, а **интенционал знака** – это характеристика концепта, выраженная через общие свойства денотата.

При анализе текстов используют понятия **дискурс (связный текст)** – два или более предложений, находящиеся друг с другом в смысловой связи, и **лингвистическая совместимость** – способность воспринимать и интерпретировать языковую форму представления информации.

Эти понятия будут использоваться в следующих главах.

Остается определить два важных, даже основополагающих, термина – это:

абстракция – процесс формирования образов реальности (представлений, понятий, суждений) посредством отвлечения и пополнения;

понятие – мысль, отражающая в обобщенной форме предметы и явления действительности и связи между ними посредством фиксации общих и специфических признаков, в качестве которых выступают свойства предметов и явлений и отношения между ними.

2. СТРУКТУРА ЕСТЕСТВЕННЫХ ЯЗЫКОВ

2.1. Слово в естественном языке

Человеческая речь – это, прежде всего, звуковая (устная) речь. Письменная форма речи представляет собой лишь ее обедненное графическое отображение. Тем не менее письменная речь является весьма эффективным средством человеческого общения.

Существуют различные виды письменности. Чаще всего при их создании используется фонетический принцип, когда с помощью графических символов (букв) обозначаются минимальные смысловые различительные отрезки звуковой речи (фонемы), а связная речь представляется в виде последовательности букв. На таком принципе построена и русская письменность, хотя соответствие между фонемами и буквами не всегда однозначное. Некоторые фонемы могут обозначаться различными буквами, а одни и те же буквы в различных позициях слова могут обозначать различные фонемы. Положение осложняется еще и исторической традицией, сохраняющей графические образы отрезков речи, изменивших свой первоначальный звуковой состав. В дальнейшем рассматривается только письменная форма речи.

В письменных текстах есть много условностей и элементов формализации. Например, довольно условно устанавливаются границы между словами, предложениями и другими единицами речи, а для обозначения этих границ широко применяются различного рода разделители:

- пробелы – для обозначения границ между словами;
- прописные буквы и знаки препинания – для обозначения границ между предложениями и составными частями предложений;
- абзацные отступы – для обозначения границ между связанными по смыслу группами предложений и т. п.

Слово – законченная последовательность знаков определенной длины, воспринимаемая как элемент обработки с определенным семантическим содержанием. Слово – минимальная, формально выделяемая единица связного текста, но оно – не минимальная единица смысла и может состоять из одной или нескольких *морфем*. В составе слов различают *корневые морфемы* (корни), *префиксы* (приставки) и

суффиксы. Основную смысловую нагрузку несет корень, а префиксы и суффиксы выступают в роли модификаторов смысла. Например, в слове выступающий можно выделить пять морфем: вы – ступ – а – ющ – ий.

Здесь морфема ступ – корень слова, морфема вы – префикс, морфемы а, ющ, ий – суффиксы (суффикс ий является грамматическим окончанием).

В табл. 1 приведено восемь групп однокоренных слов, у которых с помощью дефисов и пробелов обозначены границы между корнями и примыкающими к ним префиксами и суффиксами.

В этой таблице есть:

- слово пуск, состоящее только из одной корневой морфемы;
- слова, имеющие только корни и префиксы (обрыв, прорыв, обход и др.);
- слова, имеющие только корни и суффиксы (дается, пускать, строительство и др.);
- слова, имеющие и корни, и префиксы, и суффиксы (задающий, поддается и др.).

Членение слов на морфемы, равно как и членение текста на слова, словосочетания, предложения, сверхфразовые единства – дело непростое (**морфология** – учение о грамматических формах отдельных слов).

Правда, определению границ слов носители языка обучаются с первых шагов овладения письменностью, а сами слова фиксируются в нормативных словарях. Что же касается остальных единиц речи, то здесь ситуация более неопределенная. Подразделение на фонетику, морфологию и синтаксис произошло посредством дробления и механического членения. Язык изучают не в процессе его становления, а в его состоянии. Его рассматривают как нечто данное и завершенное. Живая речь разлагается на предложения, члены предложения, слова, слоги и звуки. Под **слогом** понимают часть слова, допускающую независимое обращение и обработку.

Этот метод может привести к ценным наблюдениям, но и одновременно может стать источником ошибок. Ошибки начинаются тогда, когда убеждают себя, что указанное членение находит основание в самом организме человеческой речи, что оно представляет собой нечто большее, чем абсолютно произвольное, механическое и насильственное рассечение. Многие считают, что предложение представляет естественную единицу речи, член предложения – естественную часть предложения, а слово или слог – дальнейшее естественное подразделение. Однако в анатомии, если отделить от туловища нижнюю конечность или же перепилить берцовую кость посередине – это всегда останется механическим разрушением организма, а не естественным расчленением. Единство организма заключается не в членах и суставах, его можно разрушить, но не разложить на его естественные части.

Таблица 1

Группы однокоренных слов

Да – ется	От – бир – ались
За – да – ющий	У – бир – ающийся
Под – да – ются	Вы – бир – аемый
Пере – да – ются	Об – рыв
Раз – да – ться	Про – рыв
Из – да – ние	С – рыв
Переиз – да – ние	Под – рыв
При – да – ются	Пере – рыв
По – да – вать	Пре – рыв – ает
Препе – да – вание	Беспре – рыв – но
Про – да – ваемый	Раз – рыв – ной
С – да – ется	Вз – рыв – ной
От – да – ча	Ход – овой
У – да – ться	За – ход – ить
Об – рез – анный	На – ход – ящийся
В – рез – аются	Об – ход
От – рез – аются	Необ – ход – имый
Вы – рез – ались	В – ход – ил
Пуск	Под – ход – ить
Пуск – ать	Пере – ход – ить
О – пуск – ающийся	Пре – ход – ящий
До – пуск – аемый	До – ход – ить
С – пуск – аться	По – ход – ный
Ис – пуск – ание	Про – ход – ящий
У – пуск – ают	С – ход – имость
Зна – ть	Ис – ход – ный
При – зна – ть	Вос – ход – ящий
По – зна – ние	От – ход – ящий
Опо – зна – вательный	У – ход – ил
Распо – зна – ющий	Вы – ход – ной
Со – зна – тельный	Стро – ительство
Осо – зна – ет	Над – стро – ечный
У – зна – ть	В – стро – енный
Из – бир – аться	Под – стро – йка
Под – бир – ается	Пере – стро – иться
Раз – бир – аются	До – стро – йка
Из – бир – ательность	По – стро – или
Со – бир – ающий	У – стро – йство

Анатом производит свои разрезы, конечно, не произвольно, но избирает такие места, которые представляются ему наиболее удобными. Точно так же разделение на звуки, слова, основы, суффиксы и т. д. является не наиболее естественным, а наиболее удобным.

В процессе функционирования в речи слова приобретают различные формы. Это могут быть формы *словоизменения* и *словообразования*. Граница между ними условная, и различные авторы проводят ее по-разному. Можно, например, считать, что формы склонения существительных и прилагательных, формы спряжения глаголов настоящего и будущего времени, формы изменения глаголов прошедшего времени, кратких прилагательных и кратких причастий по родам и числам являются формами словоизменения, а все остальные трансформации слов – формами словообразования.

Изменения форм слов могут носить различный характер. Они могут быть связаны как с изменением основы слова, так и с изменением его окончания. Изменение буквенного состава основ имеет место, например, в следующих парах форм слов: сижу – сидишь, шел – шли, тренировка – тренировок, нес – несли, кто – кого, время – времени, судно – суда, человек – люди.

Изменение окончаний является основным способом образования различных словоизменительных форм слов. В русском языке оно используется как самостоятельно, так и в сочетании с изменением основ слов.

В табл. 2 приведены примеры образования различных форм слов. При этом одна форма каждого слова указана полностью, а другие его формы представлены лишь своими окончаниями (например, телефон-а, телефон-у, телефон-ом...). Символ + обозначает «нулевое» окончание (отсутствие окончания).

По характеру изменения буквенного состава все основы слов могут быть отнесены к одному из следующих четырех типов:

- тип I – неизменяемые основы слов;
- тип II – основы слов, у которых имеет место чередование гласных;
- тип III – основы слов, у которых имеет место чередование согласных;
- тип IV – изменяемые основы слов, не отнесенные к типам II и III.

К основам типа IV относятся, в частности, *супплетивные* формы слов (например, следующие формы слов: кто, кого, кем, что, чего, он, ему и др.).

По способу изменения грамматических окончаний (*флексий*) и своей синтаксической функции русские слова могут быть разбиты на ряд классов, которые получили название *флексивных*. *Флексивные классы изменяемых слов выделяются на основе анализа их синтаксической функции и систем падежных, личных и родовых окончаний*. Классы *неизменяемых слов* – только по синтаксическому принципу. Список флексивных классов слов приведен в табл. 3.

Таблица 2

Примеры образования различных форм слов

Телефон	а, у, ом, е, ы, ов, ам, ами, ах
Тираж	а, у, ом, е, и, ей, ам, ами, ах
Огонь	я, ю, ем, е, и, ей, ям, ями, ях
Санаторий	я, ю, ем, и, ев, ям, ями, ях
Путь	и, ем, ей, ям, ями, ях
Глаз	а, у, ом, е, ам, ами, ах
Врач	а, у, ом, е, и, ей, ам, ами, ах
Женщина	ы, е, у, ой, +, ам, ами, ах
Переводчица	ы, е, у, ой, +, ам, ами, ах
Место	а, у, ом, е, +, ам, ами, ах
Поле	я, ю, ем, ей, ям, ями, ях
Очко	а, у, ом, е, и, ов, ам, ами, ах
Главный	ого, ому, ым, ом, ая, ой, ую, ое, ые, ых, ыми
Передний	его, ему, им, ем, яя, ей, юю, ее, ие, их, ими
Годовой	ого, ому, ым, ом, ая, ую, ое, ые, ых, ыми
Наш	его, ему, им, ем, а, ей, у, е, и, их, ими
Делаю	ешь, ет, ем, ете, ют
Строю	ишь, ит, им, ите, ят
Стучу	ишь, ит, им, ите, ат
Ехал	а, о, и
Силен	а, о, ы
Присущ	а, е, и
Два	ух, ум, умя
Двое	их, им, ими
Пять	и, ью
Столько	их, им, ими

По своей синтаксической функции изменяемые слова объединены в следующие группы:

- существительные;
- прилагательные;
- глаголы в личной форме;
- глаголы прошедшего времени, краткие прилагательные и причастия;
- количественные числительные.

Группа «существительные», в свою очередь, состоит из нескольких подгрупп, выделенных по признакам рода и одушевленности (для существительных мужского и женского рода). В каждой группе и подгруппе слова распределены по флективным классам.

Таблица 3

Флексивные классы слов

Таблица 3а

Существительные

№ п/п	Слово-представитель	Окончания 1) им.п., ед.ч. 2) тв.п., ед.ч., 3) им.п., мн.ч. 4) род.п., мн.ч.
<i>Существительные мужского рода неодушевленные</i>		
001	телефон	+, ом, ы, ов
002	тираж	+, ом, и, ей
003	огонь	ь, ем, и, ей
004	перебой	й, ем, и, ев
005	санаторий	й, ем, и, ев
006	бланк	+, ом, и, ов
007	сапог	+, ом, и, +
010	лес	+, ом, а, ов
011	колодец	+, ем, ы, ев
012	путь (класс состоит из одного слова)	
013	край	й, ем, я, ев
014	брус	+, ом, я, ев
015	глаз	+, ом, а, +
016	зародыш	+, ем, и, ей
017	волос	+, ом, ы, +
020	лагерь	ь, ем, я, ей
<i>Существительные мужского рода одушевленные</i>		
021	кузнец	+, ом, ы, ов
022	солдат	+, ом, ы, +
023	сосед	+, ом, и, ей
024	врач	+, ом, и, ей
025	пролетарий	й, ем, и, ев
026	воробей	ей, ем, и, ев
027	конь	ь, ем, и, ей
030	учитель	й, ем, я, ей
031	сапожник	+, ом, и, ов
032	испанец	+, ем, ы, ев
033	юноша	а, ей, и, ей

Продолжение табл. 3а

034	мужчина	а, ой, ы, +
035	судья	я, ей, и, ей
036	товарищ	+, ем, и, ей
037	гражданин	+, ом, е, +
040	профессор	+, ом, а, ов
041	муж	+, ем, я, ем
042	Иванов	+, ым, ы, ых
043	сын	+, ом, я, ей
<i>Существительные женского рода одушевленные</i>		
044	женщина	а, ой, ы, +
045	переводчица	а, ей, ы, +
046	нутрия	я, ей, и, й
047	швея	я, ей, и, й
050	цапля	я, ей, и, й
051	санитарка	а, ой, и, +
053	Иванова	а, ой, ы, ых
<i>Существительные женского рода неодушевленные</i>		
054	речь	ь, ю, и, ей
055	грань	ь, ю, и, ей
056	колба	а, ой, ы, +
057	задача	а, ей, и, +
060	заготовка	а, ой, и, +
061	линия	я, ей, и, й
062	галерея	я, ей, и, й
063	земля	я, ей, и, ь
064	эскадрилья	я, ей, и, ий
065	статья	я, ей, и, ей
066	башня	я, ей, и, +
067	улица	а, ей, ы, +
<i>Существительные среднего рода</i>		
070	место	о, ом, а, +
071	облако	о, ом, а, ов
072	поле	е, ем, я, ей
073	сомнение	е, ем, я, й

Окончание табл. 3а

074	жилище	е, ем, а, +
075	перо	о, ом, я, ев
076	время	я, ем, а, +
077	побережье	е, ем, я, ий
100	колени	о, ом, и, ей
101	очко	о, ом, и, ов
102	ружье	е, ем, я, ей

Таблица 3б

Прилагательные

№ п/п	Слово-представитель	Окончания 1) им. п., муж. р., ед. ч., 2) им. п., жен. р., ед. ч., 3) род. п., муж. р., ед. ч., 4) им. п., мн. ч.
103	главный	ый, ая, ого, ые
104	передний	ий, яя, его, ие
105	хороший	ий, ая, его, ие
106	легкий	ий, ая, ого, ие
107	годовой	ой, ая, ого, ые
110	плохой	ой, ая, ого, ие
111	третий	ий, я, его, ие
112	этот, сам	+, а, ого, и
113	мой, твой, свой	й, я, его, и
114	наш, ваш	+, а, его, и
115	весь	ь, я, его, е

Таблица 3в

Глаголы в личной форме

№ п/п	Слово-представитель	Окончания 1,2,3-го лица, ед. ч., и 3-го лица мн. ч.
116	делать	ю, ешь, ет, ют
117	строить	ю, ишь, ит, ят
120	писать	у, ешь, ет, ут
121	стучать	у, ишь, ит, ат
122	бежать	у, ишь, ит, ут
123	хотеть	у, ешь, ет, ят
124	зависеть	у, ишь, ит, ят

Таблица 3г

**Глаголы прошедшего времени,
краткие прилагательные, причастия**

№ п/п	Слово-представитель	Окончания ед.и мн.ч.
125	ехал (глагол)	+, о, а, и
126	силен (прилагательные)	+, о, а, ы
127	присущ (прилагательные)	+, е, а, и
130	краток (прилагательные)	

Таблица 3д

Количественные числительные

№ п/п	Слово-представитель
131	два, две
132	три
133	четыре
134	двое, трое
135	четверо, пятеро и т.д.
136	прочие количественные числительные (<i>пять, шесть, семь</i> и др., изменяющиеся как слово <i>мишень</i>)
137	столько, сколько
140	оба, обе

Таблица 3е

Неизменяемые слова

№ п/п	Наименование класса слов
143	Модальные слова, неизменяемые глаголы
144	Неопределенная форма глагола
145	Неизменяемые существительные мужского рода
146	Неизменяемые существительные женского рода
147	Неизменяемые существительные среднего рода
150	Неизменяемые существительные множественного числа
151	Неизменяемые прилагательные

Окончание табл. 3е

152	Деепричастие, наречие, сравнительная степень прилагательного
153	Союзы
154	Частицы, вводные слова, междометия
155	Предлог (род. п.)
156	Предлог (дат. п.)
157	Предлог (вин. п.)
160	Предлог (тв. п.)
161	Предлог (предл. п.)
162	Предлог (род., тв. п.)
163	Предлог (вин., тв. п.)
164	Предлог (вин., предл. п.)

Определение принадлежности изменяемого слова к синтаксической группе или подгруппе обычно не вызывает затруднений, так как в основу принятого здесь разделения на группы и подгруппы положена традиционная классификация слов. Следует лишь учитывать, что, наряду с полными прилагательными, к группе «прилагательные» отнесены также полные причастия, порядковые числительные, а также количественное числительное «один». При выделении окончания слова возвратные частицы *ся*, *сь* и внутренний мягкий знак (мягкий знак, стоящий между основой и ненулевым окончанием слова) опускаются. Список различных окончаний слов приведен в табл. 4.

Для характеристики системы окончаний слова нет необходимости перечислять окончания всех его форм. Обычно достаточно сделать это лишь для нескольких типичных форм. В качестве таких типичных форм для группы:

«существительные» приняты формы именительного и творительного падежей единственного числа и именительного и родительного падежей множественного числа;

«прилагательные» – формы именительного падежа единственного числа мужского и женского рода, родительного падежа единственного числа мужского рода и именительного падежа множественного числа;

«глаголы в личной форме» – формы первого, второго и третьего лица единственного числа и третьего лица множественного числа.

В группе «глаголы прошедшего времени, краткие прилагательные и причастия» окончания указаны для всех форм единственного и множественного числа. Здесь флексивный класс определяется с помощью системы окончаний и указания на принадлежность к одной из частей речи (глагол, причастие, прилагательное). Флексивные классы группы «количественные числительные» характеризуются только словами-представителями.

Таблица 4

Список окончаний слов

1 – ами	21 – ат	41 – мя	61 – ям
2 – его	22 – ах	42 – ов	62 – ят
3 – еми	23 – ая	43 – ое	63 – ях
4 – ему	24 – ев	44 – ой	64 – яя
5 – емя	25 – ее	45 – ом	65 – +(нуль)
6 – ете	26 – ей	46 – ою	66 – а
7 – ешь	27 – ем	47 – ум	67 – е
10 – ими	30 – ет	50 – ут	70 – и
11 – ите	31 – ех	51 – ух	71 – й
12 – ишь	32 – ею	52 – ую	72 – о
13 – ого	33 – ие	53 – ые	73 – у
14 – ому	34 – ий	54 – ые	74 – ы
15 – умя	35 – им	55 – ым	75 – ь
16 – ыми	36 – ит	56 – ых	76 – ю
17 – ями	37 – их	57 – ют	77 – я
20 – ам	40 – ми	60 – юю	

Некоторые классы существительных мужского и женского рода имеют одинаковые окончания во всех формах, принятых в качестве типичных, хотя другие их формы не совпадают. Иллюстрацией этому могут служить пары слов: огонь – путь, перебой – санаторий, сосед – врач, нутрия – швея, грань – речь, линия – галерея. Дополнительным признаком, необходимым для различения классов, здесь может служить информация о конечной букве основы слова, а для классов со словами-представителями огонь и путь – указание на то, что слово “путь” является единственным представителем класса (табл. 3).

В русском языке имеет место сильная корреляция между грамматической информацией к словам и буквенным оформлением их концов. Это легко обнаружить с помощью так называемых *обратных словарей*. Элементы таких словарей располагаются не в обычном лексикографическом порядке, а в обратном – так, что одинаковые концы различных слов оказываются стоящими рядом. Если теперь назначить всем словам индексы (номера) их флексивных классов, то окажется, что, как правило, одинаковым рядом стоящим концам слов будут соответствовать и одинаковые флексивные классы. Фрагменты обратного словаря словоформ с назначенными флексивными классами приведены в табл. 5.

Следовательно, “новым” словам (словам, отсутствующим в словаре) флексивные классы могут назначаться по аналогии со словами, имеющимися в словаре, если буквенный состав их концов совпадает с

Таблица 5

Фрагменты обратного словаря словоформ

масштаба – 001	плавкие – 006
хлеба – 001	легкие – 006
служба – 056	редкие – 006
дружба – 056	жидкие – 006
перегиба – 001	далекие – 006
столба – 001	резкие – 006
бомба – 056	низкие – 006
оба – 140	узкие – 006
желоба – 001	вязкие – 006
короба – 001	великие – 006
.....
сперва – 052	устанавливали – 125
битва – 056	усиливали – 125
удобства – 070	оценивали – 125
рыболовства – 070	приваривали – 125
чувства – 070	наращивали – 125
средства – 070	требовали – 125
радиосредства – 070	потребовали – 125
производства – 070	участвовали – 125
делопроизводства – 070	способствовали – 125
шелководства – 070	чувствовали – 125
.....
фабрика – 060	наилучшей – 105
Америка – 060	вкладышей – 016
метрика – 060	зародышей – 016
Африка – 060	наибольшей – 105
Мексика – 060	меньшей – 105
тросика – 060	лежащей – 105
синтагматика – 060	служашей – 105
парадигматика – 060	общей – 105
проблематика – 060	всеобщей – 105
тематика – 060	бегущей – 105
.....
перекосов – 001	район – 001
откосов – 001	закон – 001
взносков – 001	балкон – 001
запросов – 001	окон – 070
вопросов – 001	волокон – 070
.....

Окончание табл. 5

измеренного – 103	возникнуть – 144
расширенного – 103	проникнуть – 144
растворенного – 103	крикнуть – 144
подветренного – 103	подчеркнуть – 144
рассмотренного – 103	привыкнуть – 144
предусмотренного – 103	тронуть – 144
занесенного – 103	вернуть – 144
перенесенного – 103	развернуть – 144
захваченного – 103	повернуть – 144
охваченного – 103	сунуть – 144
.....
неожиданно – 152	расписанию – 073
безнаказанно – 152	возрастанию – 073
странно – 152	отрицанию – 073
особенно – 152	возникновению – 073
мгновенно – 152	проникновению – 073
откровенно – 152	выпадению – 073
косвенно – 152	ведению – 073
собственно – 152	введению – 073
явственно – 152	возведению – 073
непосредственно – 152	приведению – 073
.....
исключает – 116	степенью – 055
ухудшает – 116	ступенью – 055
мешает – 116	осенью – 055
решает – 116	болезнью – 055
разрешает – 116	жизнью – 055
лишает – 116	цепью – 055
завершает – 116	дверью – 055
совершает – 116	смесью – 055
улучшает – 116	огнесмесью – 055
превышает – 116	статью – 065
.....
газеты – 056	острая – 103
макеты – 001	быстрая – 103
ракеты – 056	косая – 107
самолеты – 001	начатая – 103
пистолеты – 001	коробчатая – 103
.....

буквенным составом словарных слов. Для этой цели нужно выбирать такие слова из словаря, концы которых в максимальной степени совпадают с концами «новых» слов. Исследования показывают, что таким образом можно правильно назначать флективные классы слов с вероятностью 0,9. С высокой степенью вероятности можно назначать «новым» словам также и другую грамматическую информацию – принадлежность к части речи, признаки рода, лица, числа, падежа и т. п.

Наряду с рассмотренными выше способами варьирования форм слов, которые мы назвали способами словоизменения, в практике речевого общения широко используются и способы словообразования. Словообразовательные трансформации слов связаны, прежде всего, с изменением состава их префиксов и суффиксов. При этом может иметь место также чередование гласных и согласных букв в корневых морфемах (например, у пар слов проводить – проведение, относиться – отношение, убедившийся – убежденный, проносу – пронесли). Перечень наиболее часто встречающихся префиксов и сочетаний префиксов приведен в табл. 6, а перечень наиболее часто встречающихся суффиксов и сочетаний суффиксов – в табл. 7.

Таблица 6

Наиболее часто встречающиеся префиксы

БЕЗ	– беззащитный, безусловный
БЕС	– бесконечный, бесполезный
В	– введение, включить, внести
ВЗ	– взгляд, взлет, взлом
ВНЕ	– внеочередной, внеплановый
ВО	– вовлечение, вопрос
ВОЗ	– воздействовать, возложить, возрасти
ВОС	– воспроизводить, восстановление
ВС	– вскрыть, всплеск, всхолмленный
ВЫ	– выбрать, вывести, выдвинуть
ДЕЗ	– дезинформация, дезорганизация
ДЕ	– декодировать, дестабилизация
ДО	– доведение, допускать, достроить
ЗА	– завершение, заглушить, задолго, затраты
ИЗ	– изготовить, излагать, измерить
ИС	– использование, истечение, исход
МЕЖ	– межведомственный, межгосударственный
МЕЖДУ	– международный
НА	– наведение, нагревание, наземный
НАД	– надводный, надстройка, надклассовый
НЕ	– небольшой, невозможный, неподвижный

Окончание табл. 6

НИ	– никакой, ничего, ничем
О	– оказать, охарактеризовать, окончить
ОБ	– обгонять, обвал, обновить, обучение
ОБЕЗ	– обезвреживать
ОВЕС	– обескровить
ОБО	– обошли, обозначение, обозримый
ОТ	– отводить, отдать, открыть, отнести
ОТО	– отошел, отомрет, отозван
ПЕРЕ	– перевод, перегрузка, передать
ПО	– повести, подать, показать
ПОД	– подвесить, подвести, подготовить
ПОДО	– подобрать, подогреть, подошел
ПРЕ	– превращать, превысить, прекрасный
ПРЕД	– предвидеть, предлагать, предсказание
ПРИ	– прибегать, прибыть, привести, пригодный
ПРО	– проанализировать, провести, проход
ПРОТИВО	– противодействовать, противоречие
РАЗ	– разбить, развить, разговор, разделить
РАЗО	– разобрать, разорвать, разослать
РАС	– раскрывать, расход
РЕ	– реконструкция, ремонтировать, реорганизация
С	– сбить, сбросить, сдавать
СО	– собирать, совершить, содержать, соединение
У	– увидеть, удаление, удержать, укладка, укрытый

В табл. 7 каждому суффиксу и сочетанию суффиксов поставлен в соответствие номер флективного класса. При этом суффиксы (сочетания суффиксов), имеющие одинаковый буквенный состав, но относящиеся к различным флективным классам (совместимые с различными наборами окончаний), считаются разными. Например, суффиксы *н* в словах “главный” и “отрывной” и суффиксы *ов* в словах “портовый” и “годовой” – разные суффиксы.

Замена у слов одних префиксов на другие приводит, как правило, к изменению их смысла, тогда как замена суффиксов в основном связана с изменением синтаксической функции. Поэтому закономерности суффиксального словообразования играют важную роль при автоматическом распознавании смыслового тождества слов и словосочетаний.

В табл. 8 даны примеры близких по смыслу слов, отличающихся друг от друга составом суффиксов.

Таблица 7

Наиболее часто встречающиеся суффиксы и их сочетания

Буквенные коды	Флексивные классы	Буквенные коды	Флексивные классы
А	116	ЛЕНИ	073
АЕМ	103	ЛЕНН	103
АЛ	125	ЛЯ	116
АНИ	073	Н	103
АНН	103	Н	107
АТЬ	144	НИ	073
АЦИ	061	НОСТ	055
АЮЩ	105	О	152
ВШ	105	ОВ	103
ЕН	126	ОВ	107
ЕНИ	073	ОВАЛ	125
ЕНН	103	ОВАН	126
И	073	ОВАНИ	073
ИВА	116	ОВАНН	103
ИВШ	105	ОВАТЬ	144
ИЛ	125	ОНН	103
ИМ	103	ОСТ	055
ИРОВАНИ	073	СК	106
ИРУ	116	ТЬ	144
ИРУЮЩ	105	У	116
ИТЕЛЬН	103	УЮЩ	105
ИТЬ	144	ЫВА	116
ИЧЕСК	106	ЬН	103
К	060	ЮЩ	105
Л	125	Я	116
		ЯЩ	105
		ЯЮЩ	105

С целью изучения закономерностей суффиксального словообразования проводились исследования словообразовательных классов слов. Классы выявлялись путем анализа словаря, составленного по научно-техническим текстам общей протяженностью более двух миллионов слов. Среди них тексты широкой тематики занимали объем около 500 000 слов, а более полутора миллиона слов – тексты реферативных журналов по информатике, вычислительной технике и радиоэлектронике (примерно в равной доле).

Из текстов широкого профиля было отобрано около 19 000 наиболее часто встречающихся словоформ, а из текстов рефератов – около

Таблица 8

Примеры словообразовательных парадигм

Звучит – звуча, звучащий, звучаши, звучал, звучание, звучать, звучащий, звучен, звучный, звучно, звучность.
Отрыв – отрывает, отрывавший, отрывавши, отрываемый, отрывал, отрываю, отрывать, отрывающий, отрывая, отрывной.
Смешает – смешав, смешавший, смешавши, смешал, смешаны, смешанный, смешать, смешение, смешивает, смешивавший, смешиваемый, смешивал, смешивание, смешивать, смешивающий, смешивая.
Рассказ – рассказав, рассказавший, рассказавши, рассказы, рассказанный, рассказал, рассказать, рассказывает, рассказывавший, рассказываемый, рассказывал, рассказывать, рассказывающий, рассказывая, рассказчик, рассказчица.
Издает – издав, издававший, издаваемый, издавал, издавать, издавая, издавший, издадут, издал, изданы, издание, изданный, издатель, издать, издающий.
Тип – типизация, типизировал, типизированы, типизированный, типизировать, типизирует, типизируемый, типизирующий, типизируя, типичен, типический, типичный, типичны, типичность, типовой.
Организует – организация, организационный, организовав, организовавший, организовавши, организовал, организован, организованный, организовать, организовывал, организовывать, организует, организуемый, организующий, организуя.
Плоский – плоско, плоскость, плоскостной.
Очаг – очаговый.
Завод – заводской.
Сила – силовой, сильный, сильны, сильнее, сильнейший.
Европа – европейский.

9000 слов. При этом каждое слово было представлено своей наиболее частой формой. Словарь был также пополнен одной тысячей слов из частотного словаря, составленного по современным русским художественным, научно-публицистическим и деловым текстам. Словарь был создан на основе статистического анализа текстов общей протяженностью более трех миллионов слов и включал в свой состав около 29 000 словоформ. При этом в словаре оказалось около 17 000 словоизменительных основ слов и около 10 000 словообразовательных основ. Под словообразовательной основой слова понималась начальная часть его буквенного кода, остающаяся после отсечения максимального числа суффиксов и удовлетворяющая условию продуктивности. *Условие продуктивности* формулируется как способность выделенной основы образовывать осмысленные слова в сочетании с другими суффиксами.

Важной характеристикой словаря, определяющей его представительность, является полнота покрытия им текстов, по которым он не составлялся. Для оценки этой характеристики были взяты тексты рефератов по информатике, вычислительной технике, электронике и газетный текст. Отождествление слов текстов со словами словаря производилось тремя способами на основе:

- полного совпадения буквенных кодов словоформ;
- совпадения буквенных кодов слов с точностью до словоизменения (отличие допускалось только в окончаниях);
- совпадения слов с точностью до словообразования (отличие допускалось только в суффиксах и окончаниях).

Результаты эксперимента сведены в табл. 9.

Таблица 9

Полнота покрытия словарем текстов различной тематики

Характер текста	Полнота покрытия в %		
	При совпадении словоформ	При совпадении словоизменительных основ слов	При совпадении словообразовательных основ слов
Информатика	77,9	95,7	96,7
Вычислительная техника	78,4	97,2	97,8
Электроника	71,6	94,3	95,8
Геофизика	68,1	86,3	90,7
Общественно-политический текст	69,0	85,0	91,4

Из таблицы видно, что в случае отождествления слов путем их словообразовательного анализа полнота покрытия текстов по информатике, вычислительной технике и электронике колеблется в пределах от 96% до 98%, а по геофизике и газетным текстам она равна примерно 91%.

Словообразовательный класс слова может быть охарактеризован перечнем суффиксов (сочетаний суффиксов), совместимых с его словообразовательной основой. При этом два слова относятся к различным классам, если перечни суффиксов (сочетаний суффиксов), совместимых с их словообразовательными основами, отличаются друг от друга хотя бы одним элементом. Так, слова “печатать” и “наблюдать” со словообразовательными основами “печат” и “наблюд” относятся к различным классам, так как основа “печат” несовместима с суффиксом *ени*, который совместим с основой “наблюд”, а основа “наблюд”, в свою очередь, несовместима с суффиксом *ани*, совместимым с основой “печат”. Всего в научно-техническом словаре было выявлено 1126 различных словообразовательных классов слов. Длина соответствующих им списков суффиксов и сочетаний суффиксов колебалась в пределах от 2-х до 38-ми и в среднем составляла 11,7.

В процессе анализа научно-технического словаря в нем было обнаружено 669 различных суффиксов и сочетаний суффиксов. Распределены они весьма неравномерно. Так, 38 наиболее продуктивных из них встречаются у 60% слов, 72 – у 75%, 181 – у 92%, а на остальные 488 суффиксов и сочетаний суффиксов приходится только 8% слов. При этом в расчет принимались только те слова из словаря, которые имели в своем составе ненулевые суффиксы.

Русские слова могут иметь в своем составе не одну, а две, три и более корневых морфем, например:

- двухкоренные слова – пустотелый, газоразрядный, двухсекционный, дисковод, псевдокоманда;
- трехкоренные слова – сверхбыстродействующий, самолетостроительный, фотополупроводниковый, светодальномер, электрофотография;
- четырехкоренные слова – водородно-кислородный, хлорметилполистирол.

В рассматриваемом нами научно-техническом словаре однокоренных слов оказалось 80,6%, двухкоренных слов – 18,7%, а трех- и четырехкоренных слов – 0,7%.

Некоторые сложные (многокоренные) слова могут иметь внутреннюю флексию (например, завод-изготовитель, слесарь-инструментальщик), и в процессе их функционирования внутренние флексии изменяются по тем же правилам, что и флексии, стоящие в конце слова (заводом-изготовителем, слесарем-инструментальщиком). Сложные слова занимают промежуточное положение между однокоренными словами и словосочетаниями.

2.2. Словосочетание в естественном языке

В автоматизированных информационных системах, основанных на формализованной записи, используются понятия, выраженные именными словосочетаниями. **Словосочетание** – это смысловое и грамматическое объединение нескольких полнозначных слов.

Эти понятия могут обозначать различного рода объекты, их признаки, значения признаков, рубрики классификационных схем и т.п. В именных словосочетаниях главным словом (*основным носителем смысла*) является, как правило, первое слева существительное, а остальные слова служат для уточнения значения главного слова.

Именные словосочетания могут включать в свой состав следующие классы слов:

- существительные (С);
- прилагательные (П);
- предлоги (Р);
- сочинительные союзы (&);
- наречия (Н).

Наряду с полными буквенными кодами слов в составе именных

словосочетаний встречаются аббревиатуры, буквенно-цифровые обозначения и числа. Эти элементы словосочетаний обычно выступают в роли существительных и значительно реже в роли прилагательных (например, порядковые числительные в цифровом выражении).

Количество слов в наименованиях понятий колеблется в пределах от одного до десяти – пятнадцати и в среднем равно примерно трем. Слова могут находиться в различной связи друг с другом. Наиболее типичными видами связи являются связь согласования между существительными и определяющими их прилагательными, а также предложные и беспредложные связи между существительными.

Прилагательное, как правило, согласуется с существительным, к которому оно относится, в роде, числе и падеже. Существительное, выступающее в роли определения к другому существительному, располагается справа от последнего и может иметь форму родительного, творительного или, значительно реже, дательного падежа. В случае предложного управления форма существительного, стоящего справа от предлога, зависит от вида последнего.

Примеры различных структур именных словосочетаний приведены в табл. 10. Здесь каждому слову наименования понятия поставлен в соответствие символ синтаксического класса. Стрелками указано направление связей между существительными, существительными и предлогами, а также между существительными и определяющими их прилагательными, если последние расположены справа от существительных. Если прилагательные располагаются слева от определяемых ими существительных, то стрелки не ставятся. В нижних индексах символов существительных, не являющихся главными словами, указаны падежи. Падежи обозначены начальными буквами их наименований.

Наименования одних и тех же понятий могут встречаться в АИС в различной форме. Их трансформации могут быть связаны с изменением порядка следования слов, с изменением форм слов, с переходом слов из одних синтаксических классов в другие. Например, словосочетания:

- “управляемые реактивные снаряды” и “реактивные управляемые снаряды” отличаются только порядком следования слов;
- “порождающие грамматики” и “порождающая грамматика” – только формами слов;
- “заводской коллектив” и “коллектив завода” – принадлежностью определителей слова “коллектив” к различным частям речи.

Чаще всего у именных словосочетаний в различных контекстных окружениях изменяется только форма главного слова и определяющих его прилагательных (“информационно-поисковые языки дескрипторного типа” – “информационно-поисковых языков дескрипторного типа”). Но в некоторых случаях имеет место зависимость форм несогласованных определений и относящихся к ним прилагательных от числа главного слова (например, в словосочетаниях “директор автомобильного завода” – “директора автомобильных заводов”, “начальник цеха – начальники цехов”).

Таблица 10

Структурные формулы словосочетаний

№ п/п	Структурная формула	Словосочетание-представитель
1	ПС	индикаторное устройство
2	ППС	цветное индикаторное устройство
3	ПППС	управляющая цифровая вычислительная машина
4	$C \rightarrow C_P$	испытания машин
5	$C \rightarrow PC_P$	испытание электронного оборудования
6	$C \rightarrow PPC_P$	использование цифровых вычислительных машин
7	$C \rightarrow PPPC_P$	использование управляющих цифровых вычислительных машин
8	$PC \rightarrow C_P$	автоматический поиск информации
9	$PC \rightarrow C_P \rightarrow C_P$	автоматизированная система поиска и информации
10	$PPC \rightarrow PC_P$	международная автоматическая система телефонной связи
11	$C \rightarrow C_P \rightarrow C_P$	автоматизация процессов управления
12	$C \rightarrow C_P \rightarrow C_P \rightarrow C_P$	проектирование систем обработки информации
13	$PC \rightarrow C_P \rightarrow PC_T$	автоматизированная система управления воздушным движением
14	$PC \rightarrow P \rightarrow PC_P$	информационная система для административного руководства
15	$PC \rightarrow P \rightarrow C_P \rightarrow C_P$	символические языки для поиска информации
16	$C \rightarrow C_{P\&} \rightarrow C_P$	система хранения и поиска информации
17	$C \rightarrow P \rightarrow C_P \rightarrow C_P$	сопротивление в месте повреждения
18	$PC \rightarrow P \rightarrow C_T \rightarrow C_P \rightarrow P \rightarrow C_B$	электрическая сеть с возвратом тока через землю
19	$PC \rightarrow C_{P\&} \rightarrow C_P$	комбинированный трансформатор тока и напряжения
20	$C \rightarrow PP$	медь листовая красная

Опираясь на структуру именных словосочетаний, можно формальным образом устанавливать между ними различные смысловые отношения:

- отношения эквивалентности (синонимии);
- родовидовые и ассоциативные отношения.

Так, словосочетания, у которых совпадают словоизменительные основы главных слов и словообразовательные основы их определителей, оказываются, как правило, синонимами. Более точно отношения синонимии устанавливаются, если кроме совпадения указанных элементов совпадают еще и схемы синтаксических связей между ними.

Родовидовые отношения между словосочетаниями имеют место, когда совпадают словоизменительные основы их главных слов, а словообразовательные основы определителей главного слова одного из словосочетаний содержатся среди словообразовательных основ определителей главного слова другого словосочетания. При этом словосочетание с меньшим числом слов выражает *родовое* понятие, а с большим – *видовое*. Точность установления родовидовых отношений увеличивается, если схема синтаксических связей между словами в словосочетании с более широким смыслом совпадает со схемой связей между соответствующими словами в словосочетании с более узким смыслом.

Родовидовые отношения имеют место, например, между следующими словосочетаниями:

- учебные заведения – высшие учебные заведения;
- системы поиска документов – автоматизированные документальные поисковые системы;
- каналы связи – каналы телеграфной связи.

Если словосочетания не являются синонимами и не находятся в родовидовых отношениях, но тем не менее имеют одинаковый или частично совпадающий состав словообразовательных основ слов, то они, как правило, связаны и по смыслу (находятся в *ассоциативных* отношениях). Ассоциативные отношения между словосочетаниями иногда полезны при поиске информации.

Полнота установления родовидовых связей между словосочетаниями может быть существенно увеличена, если при анализе их структуры учитывать родовидовые отношения между составляющими их словами. Если, например, известно, что понятия “сортировка” и “кодирование” являются видовыми по отношению к родовому понятию “обработка”, а понятие “сообщение” – видовым по отношению к понятию “информация”, то, заменяя в словосочетании “обработка информации” исходные слова на слова, выражающие соответствующие видовые понятия, получим ряд новых, более узких по смыслу словосочетаний:

- сортировка информации;
- кодирование информации;
- обработка сообщений;
- сортировка сообщений;

– кодирование сообщений.

При установлении родовидовых связей между словосочетаниями на основе смысловых связей между составляющими их словами должны выполняться три условия:

- главному слову родового словосочетания должно соответствовать эквивалентное или более узкое по смыслу главное слово видового словосочетания;
- каждому определителю главного слова родового словосочетания должен соответствовать эквивалентный или более узкий по смыслу определитель главного слова в видовом словосочетании;
- схемы синтаксических связей между связями в родовом словосочетании и соответствующими им словами в видовом словосочетании должны совпадать.

В заключение определим основные виды сочетаемости слов:

морфо-синтаксическая сочетаемость слов – информация о части речи или синтаксическом статусе и грамматической форме слова В, которое сочетается со словом А;

лексическая сочетаемость слов – информация о том, каким должно быть слово В (или набор слов), которое синтаксически сочетается со словом А;

семантическая сочетаемость слов – информация о том, какими семантическими признаками должно обладать слово В, которое синтаксически сочетается со словом А.

2.3. Предложение

Предложение – базовая единица языка, обладающая определенной для данного языка синтаксической и смысловой законченностью.

Различают предложения простые и сложные. Обычно простое предложение состоит из группы подлежащего и группы сказуемого. Группа подлежащего обозначает предмет высказывания (то, о чем говорится), группа сказуемого – признак предмета (то, что говорится). В общем случае группа подлежащего выражается именным словосочетанием, а группа сказуемого – глаголом, кратким прилагательным или кратким причастием с определяющими или дополняющими их словами и словосочетаниями. Она может также выражаться и другими частями речи.

Приведем несколько примеров простых предложений, обозначив границу между группой подлежащего и группой сказуемого знаком – (дефис):

“Изделия из порошков – позволяют регулировать пористость в сплавах”;

“Резервы экономии – не исчерпаны”;

“Коллектив Института химии Уральского центра Академии наук – ведет работы по созданию твердых сплавов для режущего инструмента и технологической оснастки”;

“Он – сильный”;

“Эта работа – выполнена за очень короткий срок”.

В составе группы сказуемого можно различать:

- собственно сказуемое (глагол, краткое причастие, краткое прилагательное, и др.);
- дополнения (прямые и косвенные);
- обстоятельства (места, времени, цели, причины, образа действия и др.).

Таким образом, предложение представляется состоящим из ряда функциональных членов, называемых членами предложения. Члены предложения могут выражаться словами и словосочетаниями, принадлежащими к различным частям речи, но при этом наблюдается сильная корреляция между частями речи и той функциональной ролью, которую они выполняют в предложении.

Наряду с членением простого предложения на группу подлежащего и группу сказуемого, в теоретической и прикладной лингвистике часто применяется и другое его членение – на *тему* и *рему*. Это – так называемое актуальное членение. Обычно принято считать, что тема выражает известную информацию, а рема – новую. Актуальное членение предложения часто совпадает с его членением на группу подлежащего и группу сказуемого, но не всегда. Примером несовпадения актуального членения и членения на группу подлежащего и группу сказуемого может служить предложение: “Особое место в ходе обсуждения заняли вопросы повышения производительности труда”. Здесь темой предложения является словосочетание “Особое место в ходе обсуждения заняли ...” – т. е. группа сказуемого, а ремой – словосочетание “... вопросы повышения производительности труда” – т.е. группа подлежащего.

Сложное предложение может состоять из двух и более простых предложений, соединенных сочинительной или подчинительной связью. При сочинительной связи простые предложения относительно независимы друг от друга и равноправны, при подчинительной – одно из них является главным, а другие – зависимыми от него, придаточными. Придаточное предложение по сути своей является как бы развернутым членом главного предложения и выступает в той же функциональной роли, что и соответствующий член простого предложения. Так, оно может выступать в роли предложения-подлежащего, предложения-дополнения, предложения-обстоятельства и др.

В сложном предложении может быть несколько главных предложений, соединенных сочинительной связью, а каждому из главных может быть подчинено несколько придаточных предложений. Придаточные предложения, в свою очередь, тоже могут иметь подчиненные им другие придаточные предложения и т. д.

В состав простых предложений могут входить причастные и деепричастные обороты, которые по своим функциям мало чем отличаются от придаточных предложений и могут быть в них преобразованы. Например, простое предложение: “Повышенный шум

заднего моста, возникающий при движении автомобиля, может свидетельствовать о неправильной установке ведущей шестерни главной передачи” с причастным оборотом «возникающий при движении автомобиля» может быть преобразовано в сложное предложение с придаточным определительным: “Повышенный шум заднего моста, который возникает при движении автомобиля, может свидетельствовать о неправильной установке ведущей шестерни главной передачи.”

Таким образом, граница между простым и сложным предложением оказывается довольно условной. Картина осложняется еще и тем, что часто несколько простых предложений, имеющих какую-либо общую часть (группу подлежащего, группу сказуемого, дополнение, обстоятельство и т. п.), объединяются в одно предложение, в котором дублирующиеся элементы оказываются представленными только один раз, а различающиеся элементы соединяются сочинительной связью. Например, два предложения:

“Предлагаемая схема предназначена для автомобильных автоматических коробок передач”;

“Предлагаемая схема определяет моменты переключения передач с низших скоростей на высшие и обратно”;

с одной и той же группой подлежащего, но и разными группами сказуемого могут быть объединены в одно предложение:

“Предлагаемая схема предназначена для автомобильных автоматических коробок передач и определяет моменты переключения передач с низших скоростей на высшие и обратно.”

В связном тексте предложения выступают не изолированно друг от друга, а в тесной смысловой связи. В основе этой связи лежат мыслительные образы тех конкретных или абстрактных объектов (ситуаций, явлений), которые человек имеет в виду, когда он порождает текст. Образы этих объектов могут иметь определенную структуру или они структурируются человеком при их описании на естественном языке. Соответственно этому структурируется и текст.

Здесь следует напомнить, что при устном общении важную роль играют понятия **модальности** – способа понимания суждения об объекте, явлении или событии, и **суждения** – умственного акта, выражающего отношение говорящего к содержанию высказанной мысли посредством утверждения модальности сказанного и сопряженного обычно с психологическими состояниями сомнения, убежденности или веры. А предложения вначале формируются у человека как **умозаключения** – умственные действия, связующие в ряд посылки и следствия мыслей различного содержания.

В структуре мыслительных образов могут быть элементы различного масштаба, находящиеся в различных отношениях друг к другу, и эти элементы и их отношения отображаются в тексте. Возникают различные структуры текста и структурные единицы текста, получившие общее название *сверхфразовые единства*. Границы сверхфразовых единств иногда выделяются формальными графическими средствами

(абзацными отступами, обозначениями параграфов, глав и т. п.), но по большей части они распознаются только «по смыслу».

Итак, тексту соответствует некоторый мыслительный образ у его автора и порождаемый этим текстом мыслительный образ у читателя. Эти мыслительные образы могут и не совпадать, но в основе они должны быть сходными. Иначе акт коммуникации (передачи информации) с помощью текста можно считать несостоявшимся. Целью передачи информации с помощью текста является не исчерпывающее описание мыслительных образов его автора, а лишь возбуждение, создание соответствующих мыслительных образов в сознании читателя. Поэтому текст не столько «выражает», сколько «намекает», и большая часть его реального содержания оказывается «между строк». При этом автор текста обычно всегда имеет в виду определенную модель знаний своего будущего читателя.

Письменный текст, как и звуковая речь, развертывается последовательно во времени и имеет линейную структуру, тогда как мыслительные образы нелинейны. При их словесном описании может быть принят различный порядок линейной развертки, но цель описания должна быть в основном одна и та же – воссоздание в сознании читателя мыслительных образов, подобных мыслительным образам автора текста. Такое воссоздание осуществляется постепенно – путем восприятия предложения за предложением и «монтажа» возникающих при этом частичных образов в целостный мыслительный образ, соответствующий содержанию текста. При этом в каждом предложении элемент его актуального членения «тема» выполняет роль «стыковочного узла», служащего для подключения нового частичного мыслительного образа, обозначаемого этим предложением, к ранее построенному мыслительному образу.

Описанная модель восприятия текста позволяет объяснить тот факт, что связи между предложениями в нем осуществляются по большей части с помощью лексических повторов: в «стыковочных узлах» предложений повторяются наименования понятий предшествующего текста либо буквально, либо в виде синонимических – конструкций, либо в виде наименований родовых понятий и местоимений. Для связи с предыдущим текстом применяются также средства, основанные на указании его координат (слов и выражений типа «на основании вышеизложенного...», «рассмотренный нами ранее...», «описанный в главе...», «в приведенном выражении...», «здесь...», «все это...» и т. п.).

В следующем фрагменте связного текста отметим средства связи между предложениями (для удобства последующих ссылок предложения перенумерованы):

1) «Одной из наиболее перспективных разработок для сферы телекоммуникаций являются *волоконные световоды*, представляющие собой пучки очень тонких нитей из специального стекла. 2) *Они* могут передавать *лучи света* на большие расстояния по изогнутым траекториям. 3) Модулированный *луч света* может быть использован в качестве

носителя речевых, телевизионных или цифровых *сигналов*. 4) Для усиления затухающего *светового сигнала* в тракт *световода* через определенные *расстояния* включаются активные повторители. 5) Эти *расстояния* могут быть сделаны значительно большими, чем в случае медных коаксиальных *кабелей* той же пропускной способности. 6) Поэтому применение *световодов* потенциально является более дешевым методом доставки абонентам индивидуальных телевизионных программ, а также передачи телекоммуникационного трафика высокой плотности на загруженных линиях. 7) Дополнительными достоинствами *световодных кабелей* являются отсутствие помех от соседних кабелей и значительная трудность несанкционированного подключения”.

В приведенном фрагменте текста связь второго предложения с первым осуществляется с помощью местоимения *они*, соотносящегося со словосочетанием *волокнистые световоды* в первом предложении. Связь третьего предложения со вторым – с помощью словосочетания *луч света*, входящего в состав обоих предложений. Связь четвертого предложения с третьим – с помощью словосочетания *световой сигнал*, которое в данном контексте является синонимом словосочетания *луч света*. С помощью слова *световод* четвертое предложение связано также с первым и вторым предложениями. Предложения пятое и четвертое связаны между собой с помощью слова *расстояния*, а связь шестого и седьмого предложений с предыдущим текстом осуществляется через термины *световод* и *световодный кабель*.

Изучение структуры связного текста имеет огромное значение в таких областях использования информационных технологий, как:

- системы искусственного интеллекта;
- экспертные системы;
- автоматизированные системы перевода;
- общение в глобальных сетях ЭВМ и т.д.

Исследования **синтаксиса естественного языка** – системы правил соединения словоформ, словосочетаний и предложений в естественном языке – оказывают серьезное влияние на разработку **синтаксиса современных языков программирования** (набор правил, которые определяют основные внутренние структуры и последовательности символов, допустимые в языках программирования).

В исследованиях по структурам текстов особое значения имеют следующие понятия:

семантика – значения языковых единиц – слов, грамматических форм слов, словосочетаний, предложений;

семиотика – наука, изучающая знаковые системы, а также естественные и искусственные языки как знаковые системы;

метаязык – язык, используемый только для описания другого языка;

метасимвол – символ или знак, используемые для выполнения синтаксических функций, указаний и других вспомогательных целей (скобки, разделители).

3. СЕМАНТИКО-СИНТАКСИЧЕСКАЯ СТРУКТУРА И АНАЛИЗ ТЕКСТОВ

3.1. Семантико-синтаксическая структура текстов

При решении проблемы обработки текстов на естественных языках следует учитывать такие явления человеческого мышления, языка и речи, как:

возможность многоаспектного описания одних и тех же объектов и ситуаций;

различный уровень обобщения информации;

вариативность форм представления одного и того же содержания; пресуппозиции (информация, формально выраженная в сообщениях, всегда сопровождается еще и подразумеваемой информацией, в этих сообщениях формально не выраженной);

инференции (информация, являющаяся логическим следствием информации, формально выраженной в сообщениях, и той, которой располагает человек).

Проблема автоматического анализа текстов усложняется также из-за *ситуационной обусловленности* их содержания, *анафорических* (межфразовых) связей, явлений *эллипсиса* (сокращения, пропуска некоторых элементов словосочетаний и фраз), омонимии и др. Все это делает задачу автоматической обработки текстов на естественных языках чрезвычайно сложной.

С целью облегчения процесса решения этой задачи, а также в рамках теоретических исследований языковых значений был предложен ряд формализованных моделей смысловой структуры текста:

- семантические сети;
- концептуальные сети;
- фреймы и др.

По существу, эти модели отражают не только смысловую, но и синтаксическую структуру текстов. Поэтому их правильнее было бы называть семантико-синтаксическими моделями. И вообще следует заметить, что граница между семантикой и синтаксисом весьма условна, так как смысловое содержание любого отрезка текста не может быть описано без опоры на его синтаксическую структуру, а синтаксическая структура немыслима без семантического наполнения. Синтаксис и семантика так же неразрывно связаны друг с другом, как категории формы и содержания.

Семантическая сеть для описания структуры текстов определяется как множество узлов, соединенных друг с другом дугами. Каждый узел сети представляет одно понятие, смысловое содержание которого определяется совокупностью его связей с другими понятиями. Таким образом, определения понятий носят характер логического круга: содержание понятия *A* может определяться через его отношение к понятию *B*, а содержание понятия *B* – через его отношение к понятию *A*.

Каждый узел сети выступает одновременно и в качестве определяемого, и в качестве определяющего.

Различают три основных типа связей между узлами:

родовидовую связь (связь определяемого понятия с более широким по объему понятием);

определятельную связь (содержание понятия конкретизируется с помощью прилагательного или наречия);

предикативную связь, выражаемую с помощью глагола или предлога.

Каждое предложение содержит в своем составе сведения о *модальности* (время, залог, наличие или отсутствие отрицания и т. п.) и собственно *высказывание*. Высказывание состоит из глагола и его дополнений, называемых также актантами или аргументами. Актанты, в соответствии с их функциональной ролью в предложении, могут иметь следующие шесть *глубинных* (семантических) падежей:

агентивный падеж (Agentiv) – обозначает одушевленный субъект действия, выраженного глаголом;

инструментальный падеж (Instrumental) – неодушевленная сила или предмет, с помощью которого совершается действие, выраженное глаголом;

дательный падеж (Dativ) – выражает функциональную роль одушевленного существа, на которое оказывает влияние действие, выраженное глаголом;

фактивный падеж (Faktiv) – обозначает одушевленное существо или предмет, которые возникают в результате действия или состояния, выраженного глаголом;

локативный падеж (Lokativ) – обозначает местоположение или пространственные размеры действия или состояния, выраженного глаголом;

объективный падеж (Objektiv) – семантически наиболее нейтральный падеж, функциональная роль которого непосредственно определяется семантикой глагола. В ряде работ вводится еще ряд других семантических падежей (например, обозначение исходного пункта действия, конечного пункта действия и др.).

Идея семантических падежей оказала большое влияние на идеологию построения ряда других систем семантического представления текста и, в частности, на модель Симмонса. Семантическая сеть Симмонса является наиболее популярной в ряду ей подобных. Она также представляет собой сеть связанных друг с другом узлов. Но здесь связи более дифференцированы. Их можно разделить на три группы.

Первую группу составляют связи, соединяющие глагол с его актантами. Эти связи подобны *глубинным падежам* и имеют следующие обозначения:

CA1 – соответствует *агентивному падежу*;

CA2 – соответствует *инструментальному падежу*;

THEME – “тема действия”, соответствует *объективному падежу*;

SOURS – “исток действия”, обозначает исходный пункт действия или первоначального владельца некоторого объекта;

GOAL – целевой пункт действия;

LOC – место действия.

С помощью второй группы связей соединяются актанты глагола с их атрибутами. Эти связи имеют следующие обозначения:

MOD – определительная связь;

HASPART – связь типа “целое – часть”;

POSS – поссесивная связь (отношение владельца к его вещи);

ASSOC – ассоциативная связь (может обозначать различные, недостаточно строго определенные связи);

SHAPE в *SIZE* – определители формы и размеров.

Третья группа связей включает отношения, не вошедшие в первую и вторую группы. Это следующие связи:

Q – нечисловые (например, “некоторые”);

NBR – количество;

DET – артикль;

COUNT – исчисляемость;

SUP – подчиняющее понятие;

SU – подчиненное понятие;

EQ – тождество;

PARTOF – часть целого;

TOK – имя узла.

Узел в сети Симмонса характеризуется его индексом и связями с другими узлами. Индекс узла обозначается сочетанием буквы *C* и порядкового номера. Каждая связь, исходящая от узла, состоит из двух компонент: из обозначения типа связи и индекса узла, к которому она идет. В табл. 11 дан пример семантического представления фразы: “Мужчина, который поймал на удочку большую рыбу. подарил ее красивой женщине.”

Таблица 11

C1	TOK	поймать	C2	TOK	мужчина	C3	TOK	удочка
	CA1	C2		-CA1	C1		-CA2	C1
	CA2	C3		-CA1	C5			
	THEME	C4						
C4	TOK	рыба	C5	TOK	подарить	C6	TOK	большой
	SIZE	C6		CA1	C2		-SIZE	C4
	- THEME	C1		GOAL	C7			
	- THEME	C5		THEME	C4			
C7	TOK	женщина	C8	TOK	красивый			
	MOD	C8		-MOD	C7			
	- GOAL	C5						

Здесь первым узел с индексом C1 описывается четырьмя связями: связью *TOK* с именем узла *поймать*;

агентивной связью *CA1* с узлом *C2* (мужчина);
инструментальной связью *CA2* с узлом *C3* (удочка);
объективной связью *THEME* с узлом *C4* (рыба).

Для каждой из этих связей имеется инверсная связь, обозначенная в таблице знаком минус. Например, узел *C2* (мужчина) имеет две инверсных агентивных связи с глаголами *поймать* и *подарить*.

Дальнейшим шагом вперед в области формализованного описания семантико-синтаксической структуры текста является теория *концептуальных* зависимостей. Здесь, как и в падежных грамматиках и семантических сетях, также считается, что смысл предложения текста может быть отражен в одной или нескольких семантических структурах, называемых *концептуализациями*, которые состоят из ряда понятий, связанных друг с другом конечным числом семантических отношений.

Различают четыре типа понятий или концептуальных категорий:

- действие (*ACT*) – понятие, о котором можно сказать, что здесь человек или животное оказывает влияние на некоторый объект;
- источник представлений (*PP*) – понятие, характеризующее физический объект как источник представлений в сознании человека;
- модификатор представлений (*PA*) – специфицирует, уточняет понятие, являющееся источником представлений, в качестве модификаторов представлений в естественном языке выступают прилагательные;
- модификатор действий (*AA*) – определяет, модифицирует понятия, обозначающие действия (в качестве модификаторов действий часто выступают наречия).

Перечисленные четыре типа понятий являются конститuentами (составляющими элементами) двух видов концептуализации – действий и состояний. Каждая концептуализация действия имеет в своем составе ядро, состоящее из “деятеля” (*PP* – члена, который является инициатором действия) и действия (*ACT*). Эти элементы ядра связываются между собой «главной связью»:

$$PP \Leftrightarrow ACT$$

Дополнительно к обязательному для него ядру, действие может быть модифицировано другими *PP* – членами, которые могут находиться в следующих падежах:

- 1) объективный падеж $ACT \xleftarrow{O} PP$
- 2) реципиентный падеж $ACT \xleftarrow{R} \left[\begin{array}{c} \text{ } \\ \text{ } \end{array} \right] \begin{array}{c} PP \\ PP \end{array}$
- 3) директивный падеж $ACT \xleftarrow{D} \left[\begin{array}{c} \text{ } \\ \text{ } \end{array} \right] \begin{array}{c} PP \\ PP \end{array}$

Реципиентный и директивный падежи имеют по два аргумента:

“дающий – принимающий” – для реципиентного падежа и “исходный пункт – цель” для директивного.

Основное отличие теории концептуальных зависимостей от ранее рассмотренных нами семантических представлений состоит в том, что здесь все действия сводятся к ограниченному числу так называемых примитивных действий. При этом предполагается, что примитивные действия отражают некоторые психические сущности, а сложные действия можно описывать в виде структур, состоящих из примитивных действий. К числу примитивных действий относятся:

ATRANS – передача владения (контроля над) некоторой вещью от одного лица к другому (давать, брать, покупать и др.);

PTRANS – движение объекта от одного места к другому;

PROPEL – приложение физического усилия к объекту (толкать, тянуть, давить и т. д.);

MOVE – движение части тела животного, инициированное им самим;

GRASP – захват “деятелем” некоторого объекта;

INGEST – поглощение (принятие внутрь) животным некоторого объекта (есть, пить, вдыхать и т. д.);

EXPEL – выделение из животного некоторого объекта (плевать, потеть и т. д.);

MTRANS – передача информации между людьми (животными) или между различными участками их памяти (видеть, слышать, вспоминать и др.);

MBUILD – порождение новых званий на основе имеющихся (решать, умозаключать и др.);

SPEAK – порождение звуков одушевленными и неодушевленными объектами (говорить, музицировать, свистеть и др.);

ATTEND – направлять свое внимание на какой-либо объект (слушать, смотреть и др.).

Важную роль в этой модели играет инструментальное отношение. Оно применяется для соединения двух концептуализаций в одну. В отличие от падежных грамматик, где инструментальное отношение соединяет действие только с одним понятием, здесь оно может соединять концептуализации двух действий.

Наряду с действиями, в теории концептуальных зависимостей применяются еще два вида концептуализации – состояние объекта и изменение его состояния. При описании состояний объектов применяются три двухместных отношения:

POSS – владение (объект *PP1* владеет объектом *PP2*);

LOC – пространственная локализация (объект *PP1* находится на месте *PP2*);

CONT – пространственное включение (объект *PP1* находится внутри объекта *PP2*).

При этом элементы *PP* могут модифицироваться элементами *PA*. Изменение состояния объекта рассматривается как изменение значения показателя этого состояния, оцениваемого по некоторой шкале значений.

Важное место среди различных способов формализованного описания семантико-синтаксической структуры текстов занимают *фреймы* (см. 1.4.).

Мы рассмотрели ряд подходов к формализованному описанию семантико-синтаксической структуры текста, отличающихся друг от друга теми или иными чертами. Во всех случаях, по существу, используется предикатно-актантная структура. Текст представляют в виде сети, в узлах которой находятся единицы языка и речи, а эти единицы связываются друг с другом определенными отношениями. Различия в терминологии, которой пользуются авторы при изложении своих концепций, не меняют существа дела. Например, семантическая сеть – это сеть, построенная на основе ограниченного числа бинарных отношений – предикатно-актантных структур. Семантические падежи – это элементы предикатно-актантной структуры. Здесь роль имени предиката выполняет глагол, а роль его актантов – дополнения, стоящие в определенных семантических падежах. В семантической сети Симмонса в основном реализуются те же идеи, но применяется более дифференцированная система бинарных отношений.

Важно отметить следующее обстоятельство. Во всех моделях представления смысла исходят из того, что “значение” языковой единицы имеет сложную структуру. Оно описывается либо некоторым набором признаков (элементарных значений), либо совокупностью связей с другими единицами, либо сетевой структурой (например, концептуализацией).

В основе значения языковой единицы лежит тот мыслительный образ, который ассоциируется с ней в сознании человека и свойства которого проявляются во всей системе парадигматических и синтагматических отношений этой единицы с другими единицами. Число таких отношений очень велико. Оно может измеряться сотнями, тысячами и десятками тысяч. Поэтому все известные модели “смысла” языковых единиц являются довольно грубой аппроксимацией реальной картины, поскольку в них используется лишь ограниченное число отношений.

“Значения” языковых единиц нельзя исчерпывающим образом раскрыть в отрыве от процессов мышления человека, опираясь только на текст. Текст – это лишь внешнее проявление этих процессов, на основе текста можно получить только часть необходимой информации. Для получения остальной информации нужно моделировать процессы мышления человека.

3.2. Введение в семантико-синтаксический анализ текстов

В АИС семантико-синтаксический анализ текстов производится с целью формализованного представления их структуры – *выделения в них смысловых единиц и установления связей между ними*.

При этом структура текстов может интерпретироваться по-разному

и описываться на различных формализованных языках. Цели и результаты анализа могут быть разными, и термином «анализ» обозначают множество различных процедур, которые имеют между собой лишь то общее, что предложения каким-то образом расчленяются и трансформируются в другую структуру. При этом всегда необходимо уточнять, на какие составные части расчленяется предложение.

Естественными составными частями текста являются прежде всего речевые отрезки, обозначающие понятия: слова, словосочетания, фразы, сверхфразовые единства. Морфемы (корни, префиксы, суффиксы) тоже являются значащими отрезками текста, но они не обозначают понятий, если не становятся самостоятельными словами.

Границы между словами и предложениями указываются в тексте с помощью пробелов, точек и прописных букв. Что же касается остальных единиц, то их границы не всегда можно обнаружить по формальным признакам. Они «отмечены» лишь в сознании человека – теми понятиями, которые ассоциируются с ними. Содержание понятий значительно богаче содержания слов и выражений, которыми они обозначаются. Слова и выражения отражают лишь те признаки (часто не самые важные), по которым эти понятия могут быть выбраны из совокупности всех других понятий.

Наряду с признаками, выраженными в их наименованиях, понятия могут иметь множество других признаков, не получивших в этих наименованиях никакого отражения, но тем не менее оказывающих влияние на синтагматические и парадигматические отношения между ними. Многообразие признаков, характеризующих одни и те же понятия, являются объективной основой для существования различных способов их словесного описания.

Из предыдущих рассуждений следует, что для правильного анализа текстов необходимо располагать не только информацией о встречающихся в них словах и словесных выражениях, но и о понятиях, ими представляемых. И чем полнее будет эта информация, тем лучше. А еще лучше иметь *модель мыслящего субъекта*, в которой достаточно полно была бы представлена не только система понятий и их словесных обозначений, но и система знаний о соответствующей предметной области – «модель мира». Все известные способы автоматического анализа текстов еще не удовлетворяют этим требованиям и поэтому несовершенны. Но многие из них вносят существенный вклад в решение проблемы и используются при решении ряда практических задач.

Среди различных способов автоматического анализа текстов видное место занимают способы, базирующиеся на концепции порождающей трансформационной грамматики Хомского. Согласно этой концепции, каждое предложение можно рассматривать как результат некоторого *порождающего* процесса, связанного с последовательной заменой одних символов и сочетаний символов на другие. Порядок замены символов задается списком подстановок, в левой части которых стоят заменяемые символы или последовательности символов, а в правой

части – заменяющие. При этом различают *нетерминальные* и *терминальные* символы: нетерминальные символы могут заменяться на другие символы и сочетания символов, а терминальные не могут.

Исходным пунктом процесса порождения предложения является начальный символ *S*, обозначающий предложение в целом, а его конечным результатом – цепочка слов, являющихся терминальными символами. Вначале символ *S* заменяют на сочетание символов *NP* и *VP* (группа подлежащего и группа сказуемого). Далее каждый из этих символов заменяют на сочетание других нетерминальных символов, обозначающих более мелкие структурные элементы предложения, эти последние – на символы еще более мелких структурных элементов и т. д., пока в результате последовательных замен не появятся *конкретные слова* (терминальные символы). После замены всех нетерминальных символов на терминальные процесс порождения предложения заканчивается. Если в процессе порождения сохранить информацию о всех структурных элементах предложения, из которых оно составлялось, то тем самым будет получено описание его семантико-синтаксической структуры.

Наряду с подстановками, предназначенными для первоначального порождения предложений, в порождающей трансформационной грамматике могут применяться и другие подстановки, описывающие правила их трансформации. В левой части таких подстановок указывается последовательность символов, характеризующая исходную структуру, а в правой – структуру ее заменяющую.

Для анализа структуры предложений порождающая трансформационная грамматика может применяться двумя способами – “сверху – вниз” и “снизу – вверх”. По первому способу порождающая процедура функционирует в обычном порядке, но подстановки применяются таким образом, чтобы терминальные символы раньше появлялись в начале предложения. Если они совпадают со словами анализируемого предложения, то переходят к порождению соседних справа терминальных символов, если не совпадают, то ищутся другие варианты подстановок, чтобы такое совпадение имело место. Так постепенно порождаются все слова анализируемого предложения, а заодно описывается и его семантико-синтаксическая структура. При анализе “снизу – вверх” реализуется процесс, обратный процессу порождения предложения. При этом подстановки применяются в обратном порядке (справа налево), а исходная последовательность терминальных символов заменяется на символ *S*. Когда это удастся сделать, то оказывается описанной и семантико-синтаксическая структура предложения.

Процедуры анализа текстов на основе порождающей трансформационной грамматики сложны в реализации и не всегда достигают цели. Причиной последнего является то обстоятельство, что процесс анализа носит формальный характер и не опирается на содержание понятий, соответствующих структурным элементам предложений.

Более простым в реализации является анализ, основанный на

использовании сетей переходов. Сеть переходов состоит из ряда так называемых состояний, которые связаны друг с другом ориентированными, то есть проходимыми только в одном направлении, отношениями. При этом на отдельные отношения могут быть наложены условия, которыми определяется, когда каждое отношение (стрелка между узлами) может быть “пройдено” (переход может быть реализован). В процессе анализа проверяется, соответствует ли последовательность классов слов анализируемого предложения одной из последовательностей, которые могут быть “пройденны” в сети переходов.

При попытках применения рассмотренных выше способов анализа к русским текстам наряду с общими нерешенными проблемами возникают еще и дополнительные трудности, связанные с богатой системой словоизменения и словообразования в русском языке и с более свободным (по сравнению, например, с германскими и романскими языками) порядком слов в предложении. Это делает необходимым создание специальных достаточно мощных процедур морфологического анализа и учета особенностей русского синтаксиса. Далее рассматриваются способы анализа и синтеза текстов, ориентированные прежде всего на специфику русского языка.

3.3. Морфологический анализ и синтез слов

3.3.1. Общее описание морфологического анализа и синтеза

Морфологический анализ слов применяется с целью отождествления их различных форм и получения грамматической и семантической информации, необходимой на последующих этапах обработки текстов. *Морфологический синтез* – с целью получения различных форм слов при декодировании текстовой информации и выдаче ее человеку. Морфологический анализ и синтез могут строиться как на базе словаря основ слов, так и на базе словаря словоформ.

Различные способы морфологического анализа и синтеза разрабатываются для систем автоматического перевода текстовых сообщений с русского языка на иностранные и с иностранных языков на русский, а также в связи с задачами общения с автоматизированными информационными системами.

Морфологический анализ и синтез слов производится с помощью словаря основ (см. п. 3.3.2.) и ряда вспомогательных таблиц. В словарь включены основы простых и сложных слов без внутренней флексии. Для сложных слов с внутренней флексией типа “слесарь-инструментальщик”, “завод-изготовитель” и т. п. в словаре приведены лишь основы простых слов, входящих в состав этих сложных слов. Если слово имеет несколько форм основ, то в словарь, как правило, включают все формы основ слов. Исключение составляют лишь изменяемые основы типа II (см. гл. 3, основы с чередованием гласных, табл. 15), которые представлены в словаре только в одной из возможных форм. Каждой

основе словаря ставится в соответствие сочетание кода основоизменяющего класса и кода флективного класса, а омонимичной основе – серия сочетаний таких кодов.

Морфологический анализ слова начинается с его флективного анализа. *Флективный анализ* слова производится для:

- правильного выделения его основы;
- замены буквенного состава основы ее порядковым номером по словарю;
- определения грамматической информации к слову.

После флективного анализа номера основ типа III (см. гл. 3, основы с чередованием согласных, с табл. 17) заменяются номерами канонических форм основ (в частности, это может быть замена на тождественный номер, если анализируемое слово имело каноническую форму основы).

Понятия *канонической* (главной) и *вариантной* форм основы слова, а также процедуры замены вариантных форм основ на канонические потребовалось ввести в связи с необходимостью отождествлять различные формы слов на последующих этапах анализа текстов. Каноническая форма для основ типа II, III будет определена ниже.

В процессе флективного анализа основа слова может не найтись в словаре. Это возможно в тех случаях, когда:

- анализируемое слово имеет основу типа II в вариантной форме;
- анализируемое слово является сложным словом с внутренней флексией;
- основа анализируемого слова не представлена в словаре ни в канонической, ни в вариантной форме.

До окончания флективного анализа слова обычно неизвестно, какой из трех перечисленных случаев имеет место. Вначале анализируемое слово проверяется на возможность наличия вариантной формы основы типа II. Если эта возможность вероятна, то вариантная форма основы заменяется на каноническую и проверяется правильность этой замены с помощью словаря основ. При положительном результате проверки определяется номер основы и грамматической информации к слову.

Если анализируемое слово не содержит в своем составе вариантной формы основы типа II, то оно проверяется на сложность (по наличию дефиса между частями сложного слова). Сложное слово расчленяется на составляющие его простые слова, которые затем подвергаются флективному анализу.

Морфологический синтез слов в первом приближении можно рассматривать как процесс, обратный по отношению к их анализу. Однако при морфологическом синтезе не возникают трудности, аналогичные трудностям, связанным с отождествлением различных буквенных образов слов и разрешением омонимии основ слов. Кроме того, исходные данные для морфологического синтеза слов отличаются от результатов морфологического анализа тем, что здесь номер основы слова сопровождается однозначной морфологической информацией, поэтому синтез форм слов значительно проще их анализа.

Синтез форм неизменяемых слов сводится к простой выборке из словаря буквенного состава их основ. В некоторых случаях к ним присоединяется возвратная частица. Формы изменяемых слов составляются из буквенных кодов их основ и окончаний. В случае необходимости к основе слова присоединяется “внутренний” мягкий знак, а к окончанию – возвратная частица *ся* или *сь*. Кроме того, канонические формы основ типов II, III заменяются на варианты. Необходимость замены канонической формы основы на вариантную определяется по номеру основы и сопровождающей его грамматической информации.

3.3.2. Флексивный анализ и синтез

Флексивный анализ изменяемых слов производится с помощью морфологической таблицы с двумя входами. Строкам этой таблицы поставлены в соответствие порядковые номера окончаний, а столбцам – номера флексивных классов слов (см. табл. 13). На пересечении строк и столбцов морфологической таблицы для каждого фактически возможного сочетания номера флексивного класса и номера окончания изменяемого слова указывается номер соответствующей морфологической информации.

В качестве морфологической информации для синтаксического класса слов указывается:

“существительные” – число и падеж;

“прилагательные” – род, число и падеж;

“глаголы в личной форме” – число и лицо;

“глаголы прошедшего времени, краткие прилагательные и причастия” –

род и число;

“количественные числительные” – падеж.

Морфологическая информация отдельных форм слов, рассматриваемых вне контекста, обычно бывает многозначна. Поэтому им могут быть поставлены в соответствие наборы упомянутых выше морфологических характеристик. Возможные наборы морфологических характеристик для различных синтаксических классов слов сведены в табл. 12, где каждому набору присвоен определенный порядковый номер.

В табл. 12 грамматическая информация представлена в закодированном виде. Здесь используются следующие условные обозначения.

Для синтаксического класса “существительные” первая цифра в каждой паре восьмеричных цифр указывает на грамматическую категорию числа, вторая – на падеж слова. При этом цифра 1 на первом месте означает единственное число, цифра 2 – множественное число. Цифры 1, 2, 3, 4, 5, 6, стоящие на втором месте, обозначают соответственно именительный, родительный, дательный, винительный, творительный и предложный падежи. Последовательность пар восьмеричных цифр описывает случаи многозначности информации о формах слов.

Таблица 12

Грамматическая информация к словоформам
(для изменяемых слов)

№ п/п	Информация	№ п/п	Информация
<i>1. Существительные</i>		<i>2. Прилагательные</i>	
1	11	41	11, 14
2	11, 14	42	11, 14, 32, 33, 35, 36
3	11, 14, 16	43	12, 14, 22
4	11, 14, 22	44	13, 23
5	11, 22, 24	45	15, 16, 25, 26, 43
6	12	46	15, 22,
7	12, 13, 15, 16	47	15, 25, 43
10	12, 13, 16	50	16, 26
11	12, 13, 16, 21	51	21, 24
12	12, 13, 16, 21, 24	52	21, 24, 41, 44
13	12, 14	53	31
14	12, 14, 21	54	32, 33, 35, 36
15	12, 21	55	34
16	12, 21, 24	56	41, 44
17	13	57	42, 44, 46
20	13, 16	60	45
21	14	<i>3. Глаголы в личной форме</i>	
22	15	61	1
23	15, 22	62	2
24	15, 22, 24	63	3
25	15, 23	64	4
26	16	65	5
27	16, 21	66	6
30	16, 21, 24	<i>4. Глаголы прошедшего времени, краткие прилагательные</i>	
31	21	67	1
32	21, 24	70	2
33	22	71	3
34	22, 24	72	4
35	22, 24, 26	<i>5. Количественные числительные</i>	
36	23	73	1, 4
37	25	74	2, 3, 6
40	26	75	2, 4, 5
		76	3
		77	5

Для синтаксического класса «прилагательные» первая цифра в каждой паре восьмеричных цифр обозначает род и число, а вторая – падеж слова. Цифра 1 на первом месте означает, что прилагательное имеет форму мужского рода единственного числа, цифра 2 является признаком среднего рода единственного числа, цифра 3 – признаком женского рода единственного числа, цифра 4 – признаком множественного числа. Падежи прилагательных обозначаются так же, как и падежи существительных.

Морфологическая информация слов, принадлежащих к синтаксическим классам «глаголы в личной форме», «глаголы прошедшего времени, краткие прилагательные и причастия», «количественные числительные», обозначается в табл. 12 одной цифрой, а в случае многозначности – последовательностью цифр. При этом для синтаксического класса «глаголы в личной форме» цифры 1, 2, 3 обозначают первое, второе и третье лицо единственного числа, а цифры 4, 5, 6 – первое, второе и третье лицо множественного числа. Для синтаксического класса «глаголы прошедшего времени, краткие прилагательные и причастия» цифры 1, 2, 3 обозначают формы мужского, среднего и женского рода единственного числа, а цифра 4 – форму множественного числа. Формы слов синтаксического класса «количественные числительные» характеризуются только падежом, который кодируется так же, как и у существительных и прилагательных.

Один из возможных способов линейной записи морфологической таблицы (для размещения в памяти ЭВМ) иллюстрирует табл. 13 (приводится только начальный участок морфологической таблицы). Здесь каждому номеру класса (полужирные числа) поставлен в соответствие столбец пар чисел, разделенных тире. Число, стоящее в каждой паре чисел слева от тире, является номером окончания (по табл. 4), а число, стоящее справа от тире – номером морфологической информации (по табл. 12), соответствующей сочетанию номера флексивного класса и номера окончания слова. Общее количество пар чисел в табл. 13 равно количеству непустых клеток двумерной морфологической таблицы.

При известном флексивном классе и окончании слова его флексивный анализ может быть сведен к выборке информации из табл. 4, 12, 13 в следующем порядке.

Сначала по табл. 4 буквенный код окончания заменяется его номером. Затем по номеру флексивного класса и номеру окончания из табл. 13 выбирается номер морфологической информации о слове. Наконец, с помощью табл. 12 номер морфологической информации заменяется соответствующим набором морфологических характеристик.

Приведем пример флексивного анализа слов. Пусть требуется проанализировать формы слов “тираж” и “стола”, которые принадлежат к флексивным классам 002, 001 и имеют окончания + (нуль) и а.

Заменив по табл. 4 буквенные коды окончаний на их номера 65, 66, входим в табл. 13 и для сочетаний номеров классов и номеров окончаний (002, 65), (001, 66) определяем номера 02, 06 наборов

Таблица 13

Морфологическая таблица

001	002	003	004	005	006	007	010
01-37	01-37	17-37	17-37	17-37	01-37	01-37	01-37
20-36	20-36	26-33	24-33	24-33	20-36	20-36	20-36
22-40	22-40	27-22	27-22	27-22	22-40	22-40	22-40
42-33	26-33	61-36	61-36	61-36	42-33	45-22	42-33
54-22	45-22	63-40	63-40	63-40	45-22	65-04	45-22
65-02	65-02	67-26	67-26	70-30	65-02	66-06	65-02
66-06	66-06	70-32	70-32	71-02	66-06	67-26	66-16
67-26	67-26	75-02	71-02	76-17	67-26	70-32	67-26
73-17	70-32	76-17	76-17	77-06	70-32	73-17	73-17
74-32	73-17	77-06	77-06		73-17		
011	012	013	014	015	016	017	020
01-37	17-37	17-37	17-37	01-37	01-37	01-37	17-37
20-36	26-33	24-33	24-33	20-36	20-36	20-36	26-33
22-40	27-22	27-22	45-22	22-40	22-40	22-40	27-22
24-33	61-36	61-36	61-36	45-22	26-33	45-22	61-36
27-22	63-40	63-40	63-40	65-04	27-22	65-04	63-40
65-02	70-12	67-26	65-02	66-16	65-02	66-06	67-26
66-06	75-02	71-02	66-06	67-26	66-06	67-26	75-02
67-26		76-17	67-26	73-17	67-26	73-17	76-17
73-17		77-16	73-17		70-32	74-32	77-16
74-32			77-32		73-17		
021	022	023	024	025	026	027	030
01-37	01-37	17-37	01-37	17-37	17-37	17-37	17-37
20-36	20-36	26-34	20-36	24-34	24-34	26-34	26-34
22-40	22-40	45-22	22-40	27-22	26-01	27-22	27-22
42-34	45-22	61-36	26-34	61-36	27-22	61-36	61-36
45-22	65-05	63-40	45-22	63-40	61-36	63-40	63-40
65-01	66-13	65-01	65-01	70-27	63-40	67-26	67-26
63-13	67-26	66-13	66-13	71-01	67-26	70-31	75-01
67-26	63-17	67-26	67-26	76-17	70-31	75-01	76-17
73-17	74-31	70-31	70-31	77-13	76-17	76-17	77-14
74-31		73-17	73-17		77-13	77-13	

морфологической информации. По табл. 12 находим, что морфологическая информация к слову “тираж” определяется набором 11, 14 (именительный и винительный падежи единственного числа), а к слову “стола” набором 12 (родительный падеж единственного числа).

Номер флексивного класса основы определяется после ее выделения из состава анализируемого слова. Членение слова производится путем последовательного отделения его конечных букв и поиска сочетания отделенных букв, в списке окончаний. Если оказывается, что сочетание отделенных букв содержится в списке окончаний, то начальная часть слова ищется в словаре основ.

При совпадении начальной части слова с одной из основ словаря определяется номер совпавшей основы и номер ее флексивного класса или, для омонимичных основ, сочетание номеров флексивных классов. Это возможно благодаря тому, что, как указывалось выше, каждой основе словаря поставлен в соответствие номер флексивного класса, а для омонимичных основ указывается сочетание номеров флексивных классов (примером омонимичной основы является основа “осмотр”, входящая в состав форм двух различных слов – “осмотр” и “осмотреть”).

Совпадение начала слова с одной из основ словаря, а его конца с одним из окончаний возможно и при неправильном членении слова. Примером могут служить формы слов “знаков” и “управляем” с основами *знак* и *управля*. Эти формы слов могут совпасть с основами “знаков” и “управляем” слов “знаковый” и “управляемый” и неправильно расчлениться на основы *знаков* и *управляем* и нулевые окончания. Поэтому требуется проверка правильности членения слова.

Правильность членения слова определяется по морфологической таблице путем проверки найденных основы и окончания слова на совместимость. Основа и окончание слова считаются совместимыми, если клетка морфологической таблицы, соответствующая номеру флексивного класса и номеру окончания слова, не пуста (или, применительно к структуре табл. 13, если номер окончания слова содержится в левой части столбца пар чисел, соответствующего номеру флексивного класса). В противном случае основа и окончание несовместимы, и следует продолжать поиск правильного членения слова. При омонимии основ на совместимость проверяются все сочетания признаков «флексивный класс» и «окончание», полученные в результате анализа слова.

Проверка основы и окончания слова на совместимость позволяет в основном преодолеть трудности морфологического анализа, связанные с омонимией основ слов. Однако при этом остается неразрешенной такая омонимия основ слов, которая может приводить к совпадению некоторых форм различных слов. Например, у слов “техник” и “техника” совпадают несколько форм единственного и множественного числа, и вне контекста по одной форме слова нельзя определить, о каком слове идет речь. Такого рода омонимия может быть разрешена только средствами синтаксического анализа, а в некоторых случаях потребуются и семантический анализ контекста, поэтому при морфологическом анализе необходимо сохранять все возможные классы и наборы морфологической информации омонимичных словоформ.

Описанный выше процесс членения на основу и окончание

применим к словам, не имеющим в своем составе *возвратной частицы* и *мягкого знака* между основой и окончанием. Наличие одного из этих элементов несколько осложняет процесс членения слова из-за необходимости его обнаружения и выделения из состава основы или окончания. При этом обнаружение возвратной частицы *ся* или *сь* отмечается признаком возвратности, а внутренний мягкий знак исключается из состава слова.

Включение в состав слова возвратной частицы влечет за собой изменение его синтаксической роли в предложении и обычно придает ему новый смысловой оттенок (сравните слова: оборонять – обороняться, управляющий – управляющийся, пытал – пытался). Естественно поэтому рассматривать возвратную частицу как составную часть основы слова с внутренней флексией (с внутренним окончанием). Чтобы отличить основу слова с возвратной частицей от основы слова без возвратной частицы, к порядковому номеру основы, полученному по словарю, прибавляется некоторое постоянное число. Величина этого постоянного числа должна быть выбрана такой, чтобы результирующее число не совпало ни с одним номером словарной основы. С этой целью в код номера основы вносят дополнительный разряд и отмечают цифрой 1 наличие признака возвратности.

Флективный синтез изменяемых слов производится с помощью словаря основ, *обращенной морфологической матрицы* (табл. 14) и списка окончаний (табл. 4). Обращенная морфологическая таблица состоит из нескольких частей, число которых определяется количеством синтаксических классов изменяемых слов (в табл. 14 приведен фрагмент таблицы для класса существительных). По одному входу таблицы (левому) перечислены коды морфологических классов, а по другому (верхнему) – морфологическая информация (коды морфологической информации выделены). На пересечении строк и столбцов указаны номера окончаний.

При формировании буквенного кода изменяемых слов сначала номер основы заменяется ее буквенным кодом, выбранным из словаря. Затем с помощью обращенной морфологической таблицы и табл. 4 отыскивается буквенный код окончания и присоединяется к буквенному коду основы слова. В необходимых случаях к окончанию слова присоединяется также буквенный код возвратной частицы, а между основой и окончанием вставляется внутренний мягкий знак.

Поиск буквенного кода окончания проиллюстрируем на примере форм слов “столами”, “тираж”, “перебоев”, имеющих основы *стол*, *тираж*, *перебо*. Пусть для каждой формы слова указано сочетание кода флективного класса и кода однозначной морфологической информации, а последовательность этих сочетаний представлена в виде пар чисел (001, 25), (002, 11), (004, 22). Тогда, используя пары чисел в качестве исходных данных, по табл. 14 можно найти соответствующие им номера окончаний 01, 65, 24, а по табл. 4 – получить искомые буквенные коды окончаний *ами*, *+*, *ев*.

Таблица 14

Обращенная морфологическая таблица
(существительные)

11	12	13	14	15	16	21	22	23	24	25	26
001-65	66	73	65	45	67	74	42	20	74	01	22
002-65	66	73	65	45	67	70	26	20	70	01	22
003-75	77	76	75	27	67	70	26	61	70	17	63
004-71	77	76	71	27	67	70	24	61	70	17	63
005-71	77	76	71	27	70	70	24	61	70	17	63
006-65	66	73	65	45	67	70	42	20	70	01	22
007-65	66	73	65	45	67	70	65	20	70	01	22
010-65	66	73	65	45	67	66	42	20	66	01	22
011-65	66	73	65	27	67	74	24	20	74	01	22
012-75	70	70	75	27	70	70	26	61	70	17	63
013-71	77	76	71	27	67	77	24	61	77	17	63
014-65	66	73	65	45	67	77	24	61	77	17	63
015-65	66	73	65	45	67	66	65	20	66	01	22
016-65	66	73	65	27	67	70	26	20	70	01	22
017-65	66	73	65	45	67	74	65	20	74	01	22
020-75	77	76	75	27	67	77	26	61	77	17	63
021-65	66	73	66	45	67	74	42	20	42	01	22
022-65	66	73	66	45	67	74	65	20	65	01	22
023-65	66	73	66	45	67	70	26	61	26	17	63
024-65	66	73	66	45	67	70	26	20	26	01	22
025-71	77	76	77	27	70	70	24	61	24	17	63
026-26	77	76	77	27	67	70	24	61	24	17	6
027-75	77	76	77	27	67	70	26	61	26	17	363
030-75	77	76	77	27	67	77	26	61	26	17	63
031-65	66	73	66	45	67	70	42	20	42	01	22
032-65	66	73	66	27	67	74	24	20	24	01	2
033-66	70	67	77	26	67	70	26	20	26	01	222
034-66	74	67	73	44	67	74	65	20	65	01	22
035-77	70	67	76	26	67	70	26	61	26	17	63
036-65	66	73	66	26	67	70	26	20	26	01	22
037-65	66	73	66	45	67	67	65	20	65	01	22
040-65	66	73	66	45	67	66	42	20	42	01	22

Известно, что окончания прилагательных, имеющих формы винительного падежа единственного и множественного числа и согласующихся соответственно с существительными мужского и

женского рода, бывают различными в зависимости от наличия или отсутствия признака одушевленности у существительных, к которым эти прилагательные относятся. При синтаксическом синтезе в подобных случаях винительный падеж заменяется родительным, что позволяет однозначно определить окончание по обращенной морфологической таблице (это правило не распространяется на винительный падеж единственного числа прилагательных, согласованных с существительными женского рода).

Буквенный код неизменяемых слов обычно совпадает с буквенным кодом их словарных основ. Исключение составляют только слова с признаком возвратности. В последнем случае присоединяется код возвратной частицы.

Для выяснения формальных признаков, по которым можно было бы определить необходимость введения мягкого знака между основой и окончанием, был проведен соответствующий анализ частотного словаря словоформ, составленного по научно-техническим текстам. При этом проверялись две рабочие гипотезы. Первая из них заключалась в предположении, что свойство иметь внутренний мягкий знак присуще всем словам, входящим в флективные классы со словами-представителями: «брус», «воробей», «судья», «муж», «сын», «мышь», «речь», «грань», «эскадрилья», «статья», «перо», «побережье», «третий», т.е. с такими словами-представителями, которые в определенных формах могут содержать внутренний мягкий знак. Согласно второй гипотезе, предполагалось, что все слова с внутренним мягким знаком принадлежат только к одному из перечисленных выше флективных классов.

В результате анализа частотного словаря не было обнаружено ни одного примера, противоречащего этим гипотезам. Поэтому обе гипотезы могут считаться практически достоверными и использоваться при разработке алгоритмов морфологического анализа и синтеза слов.

Таким образом, для введения внутреннего мягкого знака в состав синтезируемого слова требуется, чтобы его флективный класс совпадал с одним из классов слов, допускающих эту операцию, а морфологическая информация определяла именно ту форму слова, которая у данного класса должна содержать внутренний мягкий знак. Информация о формах слов, содержащих внутренний мягкий знак, выявляется заранее и используется при составлении алгоритма морфологического синтеза.

При синтезе слов с возвратными частицами *ся* или *сь* требуется в каждом случае выяснить, какая из двух частиц должна быть выбрана. Анализ форм слов показывает, что частица *сь* обычно встречается после букв *а, е, и, о, у, ю, я* и только у инфинитива, деепричастия и у личных форм глагола. В остальных случаях употребляется частица *ся*.

3.3.3. Морфологический анализ и синтез слов с изменяемой основой

Напомним, что изменяемые основы слов бывают трех типов – II, III и IV (гл. 3). У основ слов типа II имеет место явление чередования гласных. При этом в различных формах слов заменяется или пропадает буква, предшествующая последней букве основы слова. Возможные виды чередования гласных показаны в табл. 15 и здесь же приведены примеры форм слов с основами типа II.

Таблица 15

Список подстановок для основ слов типа II

№ п/п	Класс подстановки	Вид подстановки	Примеры
1	1	о → +	Заготовок – заготовка
2	1	и → й	Достоин – достойна
3	2	е → +	Сложен – сложна
4	2	е → й	Паяк – пайка
5	2	е → ь	Колец – кольца

Основы слов типа II представлены в словаре только в канонической форме. Эта форма основы встречается в словоформах с ненулевым окончанием, отличным от мягкого знака. Вариантная форма основы бывает у словоформ с нулевым окончанием или с мягким знаком в качестве окончания, например, словоформы “колодец” и “день”. При морфологическом анализе вариантная форма основы приводится к канонической путем замены соответствующей буквы на «нуль» или на другую букву (согласно табл. 15).

Проверка основы слова на наличие беглой гласной производится после того, как основа не нашлась в словаре в результате выполнения процедуры флективного анализа. Эта проверка осуществляется только у слов, оканчивающихся на согласную или на мягкий знак. У слов, оканчивающихся на согласную, заменяется предпоследняя буква, если она является одной из букв левой части списка подстановок табл. 15. При обнаружении конечного мягкого знака он отделяется от слова (заносится вместо нулевого окончания), а затем производится замена гласной.

Подстановки табл. 15 разделяются на два класса:

класс с индексом 1 (подстановки 1 и 2),

класс с индексом 2 (подстановки 3, 4, 5).

Это разделение необходимо, чтобы обеспечить правильность морфологического анализа и синтеза слов. Индексы классов под-

становок указываются в словаре для каждой канонической формы основы типа II.

Если в анализируемом слове заменяется гласная *е*, то приходится учитывать несколько возможных вариантов замены. Для этого последовательно применяют к анализируемому слову подстановки 3, 4 и 5 табл. 15 и проверяют их на совместимость с основами словаря. Проверка на совместимость производится после отождествления трансформированной основы с одной из основ словаря. Основа словаря и подстановка считаются совместимыми, если индекс класса используемой подстановки и индекс класса подстановки, указанный в словаре, совпадают. В противном случае основа словаря и используемая подстановка несовместимы, и необходимо проверить, можно ли применить другие подстановки. Правильность применения подстановок 1 и 2 табл. 15 проверяется так же, как и в случае замены гласной *е*.

После проверки правильности замены гласной следует обычная при флексивном анализе проверка основы и окончания на совместимость и определяется номер основы и грамматической информации к слову.

Описанный порядок проверки правильности преобразования основы слова типа II в каноническую форму позволяет избежать ложных отождествлений основ слов. Действительно, сочетания индексов подстановок и букв правой части таблицы подстановок однозначно определяют беглую гласную основы исходного слова (табл. 15), а полученная гласная и неизменяемый буквенный состав словарных основ типа II полностью определяют вид основы анализируемого слова.

Для образования в процессе морфологического синтеза вариантов форм основ типа II используется табл. 16. При этом учитывается индекс класса подстановки, приписанный основе словаря, и сопровождающая номер основы грамматическая информация. К табл. 16 обращаются только тогда, когда основа слова имеет индекс класса подстановки 1 или 2, а грамматической информации соответствует окончание *+ или ь*.

Таблица 16

Список подстановок для основ слов типа II при синтезе

№ п/п	Класс подстановки	Вид подстановки	Примеры
1	1	+ → о	Кратка – краток
2	1	й → и	Достойна – достоин
3	2	+ → е	Колодца – колодец
4	2	й → е	Пайка – паек
5	2	ь → е	Льда – лед

После выборки по номеру основы ее буквенного кода он анализируется для определения вида подстановки (табл. 16). Далее производится необходимое преобразование буквенного кода основы и присоединение окончания слова.

Вид подстановки определяется по следующим правилам. Выделяется вторая от конца буква словарной основы и проверяется на совпадение с буквой *и*, если основа имеет индекс класса подстановки 1, и с буквами *и* и *ь*, если основа имеет индекс класса подстановки 2. При положительном результате проверки, в первом случае применяется подстановка 2, при отрицательном – подстановка 1. Во втором случае при положительном результате применяется подстановка 4 (если выделенная буква совпала с буквой *и*) или подстановка 5 (если выделенная буква совпала с буквой *ь*). При отрицательном результате применяется подстановка 3.

К изменяемым основам слов типа III отнесены такие основы личных форм глаголов и глаголов прошедшего времени, у которых имеет место чередование согласных. Эти основы встречаются в двух формах, отличающихся друг от друга по буквенному составу. Обе формы основы включаются в словарь. Одна из них считается канонической, другая – вариантной. У личных форм глаголов в качестве канонической принята основа формы третьего лица единственного числа, у глаголов прошедшего времени – основа формы множественного числа.

При морфологическом анализе вариантная форма основы типа III заменяется на каноническую с помощью табл. 17 по специальным признакам, внесенным в словарь основ.

Таблица 17

Список подстановок для основ слов типа III при анализе

№ п/п	Конечные буквы вариантных форм основ слов	Конечные буквы вариантных форм основ слов		Примеры
		вариант 0	вариант 1	
1	ж	д	з	Сижу – сидит, вожу – возит
2	ш	с	-	Ношу – носит
3	щ	ст	т	Очищу – очистит, сокращу – сократит
4	ч	т	-	Лечу – летит
5	г	ж	-	Могу – может
6	к	ч	-	Отсеку – отсечет
7	л	+	-	Ставлю – ставит
8	т	ч	-	Хотят – хочет
9	+	л	-	Вез – везли

Табл. 17 содержит список подстановок букв и примеры использования этих подстановок. Во втором столбце таблицы перечислены конечные буквы вариантных форм основ слов типа III, а в третьем и четвертом столбцах – конечные буквы соответствующих канонических форм. В последнем столбце приведены примеры для каждого варианта подстановок букв.

Словарными признаками, используемыми при морфологическом анализе слов с основами типа III, являются признак вида основы и признак варианта подстановки. При этом каноническая форма основы сопровождается индексом 0, а вариантная – индексом 1. Различные варианты подстановок также обозначаются индексами 0 и 1 (табл. 17).

Анализ основ слов типа III производится следующим образом. Сначала основа словаря, найденная в результате флективного анализа, проверяется на наличие признака вариантной формы. Если у основы такой признак есть, то выделяется ее последняя буква и сравнивается последовательно со всеми буквами второго столбца табл. 17 (исключая букву +). При совпадении выделенной буквы с одной из букв таблицы, она заменяется на букву (или сочетание букв) третьего или четвертого столбца в зависимости от значения признака варианта подстановки. Далее полученная основа ищется в словаре. Если трансформированная основа отождествляется с одной из основ словаря, то последняя проверяется на совместимость с окончанием и на наличие у нее признака канонической формы основы типа III. При положительном результате проверки первоначальный номер вариантной формы основы заменяется на номер ее канонической формы.

В том случае, когда выделенная буква анализируемой основы не совпадает ни с одной из букв второго столбца табл. 17, к этой основе присоединяется буква л (см. девятую строку табл. 17) и далее выполняются операции, перечисленные в предыдущем абзаце.

Формирование буквенного кода основ слов типа III при морфологическом синтезе осуществляется с помощью табл. 18 и 19. Табл. 18 служит для преобразования канонических форм основ в вариантные, а табл. 19 – для определения необходимости такого преобразования.

Структура табл. 18 аналогична структуре табл. 17. В табл. 19 перечислены различные типы распределения канонических и вариантных форм основ в зависимости от грамматической информации слов. Строкам табл. 19 поставлены в соответствие коды типов распределения, а столбцам – коды грамматической информации (см. табл. 12). На пересечении строк и столбцов указаны индексы канонических и вариантных форм основ.

Сочетание кода типа распределения и кода грамматической информации однозначно определяет необходимость введения в синтезируемое слово канонической или вариантной формы основы. Код типа распределения указывается в словаре для каждой канонической формы основы слова типа III наряду с индексом канонической формы и индексом варианта подстановки. Смысл индекса

Таблица 18

Список подстановок для основ слов типа III при синтезе

№ п/п	Конечные буквы вари- антных форм основ слов	Конечные буквы вариантных форм основ слов		Примеры
		вариант 0	вариант 1	
1	ст	щ	-	Очистит – очищу
2	т	ч	Щ	Летит – лечу, сократит – сокр ащу
3	ж	г	-	Может – могу
4	з	ж	-	Возит – вожу
5	л	ж	-	Сидит – сижу
6	с	ш	-	Носит – ношу
7	ч	к	-	Отсечет – отсеку
8	л	+	-	Везли – вез
9	+	Л	-	Ставит – ставлю

Таблица 19

**Типы распределения
канонических и вариантных форм основ слов**

Тип распределения	Грамматическая информация					
	1	2	3	4	5	6
0	0	0	0	0	0	0
1	1	0	0	0	0	0
2	0	0	0	1	1	1
3	1	0	0	0	0	1

варианта подстановки для канонических форм основ определяется табл. 18, а для вариантных – табл. 17.

Формирование буквенного кода слова начинается с выборки из словаря буквенного кода его основы. Затем по таблице 19 определяется необходимость замены канонической формы основы на вариантную. Если такой необходимости нет, то к основе присоединяется окончание. Если замена необходима, то она производится с помощью табл. 18.

Каноническая форма основы заменяется на вариантную в следующем порядке. Сначала две последние буквы основы проверяются на совпадение с сочетанием букв *ст*. Если совпадение имеет место, то эти буквы заменяют на букву *щ* (см. подстановку 1 табл. 18); если нет, то конечная буква основы отыскивается среди ненулевых букв второго

столбца табл. 18. При отождествлении конечной буквы основы с одной из букв второго столбца ее заменяют на соответствующую букву третьего или четвертого столбца (в зависимости от значения признака варианта подстановки). В противном случае к словарной основе присоединяется буква *л* (применяется подстановка 9).

Среди слов с изменяемой основой типа IV следует различать слова, способные иметь различные окончания, и слова, у которых выделять окончания трудно или практически нецелесообразно. Слова первого вида далее называются *словами с супплетивными основами*, слова второго вида – *словами с супплетивными формами*. Например, словами первого вида являются слова: *зная, время, человек, судно*, а второго вида – слова: *кто, что, чей*. Супплетивные формы основ и супплетивные формы слов заносятся в машинный словарь во всех своих вариантах и отмечаются специальным признаком, который используется при морфологическом анализе и синтезе.

Морфологический анализ слов с изменяемой основой типа IV начинается с их флективного анализа, причем слова с супплетивными формами сначала рассматриваются как неизменяемые. Далее с помощью специальных таблиц варианты формы основ заменяются каноническими, а по супплетивным формам слов выбирается соответствующая им грамматическая информация.

Процесс морфологического синтеза слов с основами типа IV состоит из двух этапов:

этапа замены канонической формы основы на вариантную, если такая замена необходима;

этапа флективного синтеза.

Необходимость выбора канонической или вариантной формы может быть определена по грамматической информации. Если грамматической информации соответствует каноническая форма основы, то следует переходить к этапу флективного синтеза; если вариантная форма основы – то исходную основу необходимо заменить вариантной.

При морфологическом анализе и синтезе супплетивные основы и супплетивные формы слов различаются номерами флективных классов (супплетивные формы слов не имеют окончаний и относятся либо к неизменяемым существительным, либо к неизменяемым прилагательным).

3.3.4. Алгоритмы морфологического анализа и синтеза

Далее приводятся обобщенные описания алгоритмов морфологического анализа и синтеза слов, построенные на основе материала, изложенного в предыдущих разделах настоящей главы. Эти описания отражают все этапы работы алгоритмов в их взаимосвязи.

Алгоритм морфологического анализа.

1. Проверка на конец текста. При положительном исходе проверки перейти к п. 2, при отрицательном – к п. 3.

2. Выход на алгоритм синтаксического анализа.
3. Занесение очередного слова в стандартное поле памяти.
4. Занесение в стандартное поле памяти номера нулевого окончания.
5. Поиск слова в словаре основ. При положительном исходе перейти к п. 6, при отрицательном – к п. 21.
6. Выборка из словаря номера основы и морфологического класса слова (или, в случае омонимии, сочетания морфологических классов).
7. Проверка выделенной основы и окончания слова на совместимость. При положительном исходе перейти к п. 8., при отрицательном – к п. 21.
8. Выборка из таблиц и запись в рабочее поле морфологической информации о слове.
9. Проверка на наличие у слова признака возвратности. При положительном исходе перейти к п. 10, при отрицательном – к п. 22.
10. Перенумеровать основу слова (занесение в код номера слова признака возвратности). Перейти к п. 22.
11. Проверка условия: «Количество оставшихся букв в слове $m = 1$ ». При положительном исходе перейти к п. 24, при отрицательном – к п. 12.
12. Отделение одной конечной буквы слова.
13. Проверка условия: «Количество отделенных букв в слове $n = 2$ ». При положительном исходе перейти к п. 18, при отрицательном – к п. 14.
14. Поиск сочетания отделенных конечных букв слова в словаре окончаний. При положительном исходе перейти к п. 15, при отрицательном – к п. 21.
15. Определение номера окончания.
16. Проверка конца слова на наличие мягкого знака. При положительном результате проверки перейти к п. 17, при отрицательном – к п. 5.
17. Отделение мягкого знака. Занесение числа 3 в счетчик отделенных букв. Перейти к п. 5.
18. Проверка отделенных букв на совпадение с частицами *ся* или *сь*. При положительном исходе перейти к п. 19, при отрицательном – к п. 14.
19. Занесение в рабочее поле признака возвратности.
20. Гашение счетчика количества отделенных букв и очистка рабочего поля, содержащего отделенные буквы. Перейти к п. 4.
21. Проверка условия: «Количество отделенных букв $n = 3$ ». При положительном исходе перейти к п. 24, при отрицательном – к п. 11.
22. Проверка на принадлежность основы слова к типу III. При положительном исходе перейти к п. 28, при отрицательном – к п. 23.
23. Проверка на принадлежность основы слова к типу IV. При положительном исходе перейти к п. 30, при отрицательном – к п. 29.
24. Проверка на принадлежность основы слова к типу II. При положительном исходе перейти к п. 31, при отрицательном – к п. 25.

25. Проверка слова на сложность. Если слово сложное, то перейти к п. 27, в противном случае – к п. 26.

26. Занесение в ответный массив признака побуквенного кодирования и буквенного кода анализируемого слова. Перейти к п. 1.

27. Членение сложного слова на составные части. Перейти к п. 3.

28. Замена вариантной формы основы типа III на каноническую (с помощью табл. 17 и словаря основ).

29. Занесение в ответный массив результатов морфологического анализа слова. Перейти к п. 1.

30. Замена вариантной формы основы типа IV на каноническую. Выборка грамматической информации для супплетивных форм слов. Перейти к п. 29.

31. Замена вариантной формы основы типа II на каноническую (с помощью табл. 19).

32. Проверка с помощью словаря основ правильности замены формы основы типа II. При положительном исходе перейти к п. 6, при отрицательном – к п. 25.

Алгоритм морфологического синтеза. Как и было сказано ранее, этот алгоритм значительно проще алгоритма анализа.

1. Проверка на конец исходного массива. При положительном исходе проверки перейти к п. 2, при отрицательном – к п. 3.

2. Конец работы алгоритма.

3. Занесение в стандартное поле памяти исходной информации об очередном слове.

4. Проверка слова на принадлежность к группе слов с изменяющимися окончаниями. При положительном исходе перейти к п. 5, при отрицательном – к п. 16.

5. Проверка основы слова на принадлежность к типу IV. При положительном исходе перейти к п. 6, при отрицательном – к п. 7.

6. Замена номера канонической формы основы типа IV на номер вариантной формы основы.

7. Выборка из словаря буквенного кода основы слова.

8. Проверка морфологической информации к слову на наличие признака чередования согласных. При положительном исходе перейти к п. 9, при отрицательном – к п. 10.

9. Замена канонической формы основы типа III на вариантную (с помощью табл. 18, 19).

10. Выявление необходимости присоединения мягкого знака к основе слова. Если такая необходимость есть, то перейти к п. 11, если нет – к п. 12.

11. Присоединение мягкого знака к основе слова.

12. Выборка с помощью табл. 4 и 14 буквенного кода окончания слова.

13. Проверка информации к слову на наличие признака чередования гласных. При положительном исходе перейти к п. 14, при отрицательном – к п. 15.

14. Замена канонической формы основы типа II на вариантную.

15. Присоединение к основе слова буквенного кода окончания. Перейти к п. 19.

16. Проверка основы слова на принадлежность к типу IV. При положительном исходе перейти к п. 17, при отрицательном – к п. 18.

17. Замена номера канонической формы основы типа IV на номер вариантной формы основы.

18. Выборка из словаря буквенного кода основы слова.

19. Проверка на наличие у слова признака возвратности. При положительном исходе перейти к п. 20, при отрицательном – к п. 1.

20. Присоединение к слову возвратной частицы. Перейти к п. 1.

Существенную часть алгоритмов морфологического анализа и синтеза слов составляют процедуры поиска в словаре. Словарь может быть оформлен различным образом и, в частности, в виде ассоциативно-адресной структуры. При этом буквенные коды основ слов должны интерпретироваться как словоформы, а сопровождающая их грамматическая информация оформляться в виде отдельного массива. Выборка грамматической информации должна осуществляться по номерам основ (точнее, по той их части, которая отражает порядок следования адресных отсылок к буквенным кодам этих основ). При наличии в словаре семантической информации она также должна выноситься в отдельный массив.

3.3.5. Сравнение различных методов анализа и синтеза

Морфологический анализ и синтез слов может производиться как на базе словаря основ слов, так и на базе словаря словоформ, поэтому представляет интерес сравнение основных количественных показателей, характеризующих эти подходы – объема словаря и времени работы алгоритмов. Объем машинного словаря зависит от многих факторов. Однако при сравнении вариантов структуры словаря прежде всего учитывают соотношение количества словарных единиц при некоторых фиксированных условиях.

В русском языке число различных словоформ значительно больше числа различных основ слов, так:

существительные могут иметь 7-10 различных форм;

полные прилагательные 10-12 форм;

глаголы настоящего и будущего времени – 6 форм;

глаголы прошедшего времени и краткие прилагательные – 4 формы

и т. д.

Если фиксировать объем словаря основ и потребовать, чтобы словарь словоформ включал все формы слов, которые могут быть образованы на базе словаря основ, то отношение числа словоформ к числу основ слов определяется выражением:

$$K = \sum_{i=1}^n M_i P_i,$$

где n – количество флективных классов слов в русском языке;
 M_i – количество попарно-различных форм у слов 1-го флективного класса,
 P_i – вероятность появления 1-го флективного класса в словаре.
 Исследования словарей показывают, что $K \approx 8$.

Однако в речевой практике не все формы слов используются в равной степени. Это приводит к тому, что при фиксированном тексте достаточно большой протяженности объем словаря словоформ оказывается примерно в два раза больше объема словаря основ (это явление наблюдалось на текстах протяженностью от 20 до 500 тыс. слов).

Время работы алгоритмов автоматического отождествления слов зависит от типа ЭВМ, которая используется для обработки текстовой информации, и от конкретной программной реализации этих алгоритмов. Имеет значение и объем словаря. Однако при прочих равных условиях *программа морфологического анализа* работает в несколько раз медленнее, чем программа отождествления слов с помощью словаря словоформ. Это обусловлено большей сложностью алгоритмов морфологического анализа и необходимостью многократного поиска по словарю при выделении основы из состава изменяемого слова.

Процедуры морфологического анализа и синтеза слов могут быть точными и приближенными. Точные процедуры основаны на использовании словарей, в которых для каждого слова указано правило изменения его формы. Эти процедуры могут применяться только к словам, которые включены в словарь. Между тем в реальных текстах всегда будут встречаться «новые» слова – слова, не содержащиеся в словаре.

«Новые» слова могут автоматически выявляться в процессе точного морфологического анализа и выдаваться на печать для ручной обработки и включения в словарь. Но такая организация работы не позволит полностью автоматизировать процессы обработки текстовой информации. Необходима процедура автоматического пополнения словарей. Это, в свою очередь, связано с необходимостью автоматического получения грамматической информации к словам.

Для анализа «новых» слов целесообразно использовать метод аналогии, основанный на связи между грамматическими признаками слов и их буквенным оформлением. Применительно к русской морфологии принцип аналогии можно было бы сформулировать следующим образом: *слова, имеющие аналогичное буквенное оформление концов, аналогичны и по грамматической информации.*

Для назначения грамматических признаков «новым» словоформам по методу аналогии необходимо иметь базовый словарь, в котором для каждой словоформы указана соответствующая ей грамматическая информация. Процедура назначения грамматических признаков выполняется в следующем порядке.

1. «Новая» словоформа сравнивается со словоформами из словаря, и фиксируются все случаи совпадения концов словоформ.

2. Из словаря выбираются словоформы, у которых длина конечных буквосочетаний, совпавших с конечным буквосочетанием «нового» слова, является максимальной.
3. Если выбирается только одна словоформа, то набор ее грамматических признаков присваивается новой словоформе.
4. Если выбирается группа словоформ, то для этой группы строится распределение частот появления различных наборов грамматических признаков и «новой» словоформе назначается наиболее частый набор.

Назначение грамматических признаков «новым» словам по методу аналогии может осуществляться и с помощью словаря основ слов. В этом случае несколько изменяется способ выбора словарных элементов, по которым производится назначение признаков. У исходной словоформы отделяются все возможные варианты грамматических окончаний, а полученные таким образом варианты основ слова сравниваются с основами словаря. В процессе сравнения фиксируются все случаи совпадения концов основ «нового» слова с концами основ из словаря при условии, что соответствующие варианты окончаний нового слова совместимы со словарными основами (совместимость основ и окончаний проверяется по табл. 13). В каждом случае определяется сумма количества совпавших букв в сравниваемых основах и количества букв в окончании «нового» слова. Из словаря выбираются основы с максимальным значением суммы. Выбранные основы используются для назначения грамматических признаков «новому» слову таким же порядком, что и словоформы из словаря словоформ. Далее у «нового» слова отделяется окончание и его основа включается в словарь вместе со своими грамматическими признаками.

3.3.6. Многоступенчатый морфологический анализ и синтез

Если строить морфологический анализ на базе словаря словоформ, то задача получения грамматических и семантических признаков для слов исходного текста сведется в основном к поиску в словаре, и только в тех случаях, когда это не удастся, придется прибегать к морфологическому анализу. При этом, чем полнее словарь, тем меньше будет удельный вес операций по анализу структуры слов и тем больше скорость обработки текстов. Кроме того, применение словаря словоформ позволяет в значительной мере преодолеть трудности, связанные с такими явлениями словоизменения и словообразования, как чередование гласных, чередование согласных и наличие супплетивных форм слов. Это достигается путем отображения в словаре парадигматических связей между словоформами независимо от их буквенного оформления.

Положительные свойства процедур морфологического анализа, построенных на базе словаря словоформ и словаря основ слов, можно сочетать в одном алгоритме. Такой алгоритм, получивший название

алгоритма многоступенчатого морфологического анализа, работает со словарем словоформ, в котором для каждой словоформы указывается длина ее словоизменительной и словообразовательной основы, а также номер ее флективного и словообразовательного класса (см. гл. 3, табл. 12).

В процессе работы алгоритма словоформы текста могут проходить три ступени анализа:

- 1) проверка на полное совпадение со словоформами словаря;
- 2) словоизменительный анализ;
- 3) словообразовательный анализ.

Наиболее простой в реализации является первая ступень анализа, наиболее сложной – третья. Чаще всего анализ словоформ ограничивается только первой ступенью, значительно реже – первой и второй ступенью, а третья ступень анализа привлекается только тогда, когда словоформы текста не удается отождествить ни с одной из словоформ словаря, ни в результате проверки на полное совпадение, ни в результате их словоизменительного анализа.

При словоизменительном анализе требуется совпадение словоизменительных основ сравниваемых слов и их принадлежность к одному и тому же флективному классу. При словообразовательном анализе – совпадение словообразовательных основ и расчленение несовпадающих частей текстовых слов на суффиксы (сочетания суффиксов) и окончания. При этом суффиксы (сочетания суффиксов) должны быть совместимы с окончаниями и со словообразовательными основами.

Совместимость суффиксов (сочетаний суффиксов) с окончаниями проверяется с помощью приписанных им номеров флективных классов, а их совместимость со словообразовательными основами – с помощью номеров словообразовательных классов. Последняя проверка осуществляется путем поиска суффикса (сочетания суффиксов) текстового слова в списке суффиксов (сочетаний суффиксов), соответствующем номеру словообразовательного класса, приписанному словарному слову. В результате морфологического анализа текстовым словам наряду с другой информацией приписываются также номера их словоизменительных и словообразовательных основ.

3.4. Анализ и синтез именных словосочетаний

3.4.1. Синтаксический анализ именных словосочетаний

Ранее указывалось, что именные словосочетания играют важную роль в системах автоматической обработки информации, так как они чаще используются для обозначения научно-технических понятий, чем однословные термины. Они могут применяться в АИС в различной форме, поэтому необходимы процедуры их морфологического, синтаксического и семантического анализа и синтеза.

В процессе синтаксического анализа наименований понятий выполняются следующие операции:

- выявляется схема связей между словами;
- каждому слову словосочетания назначается однозначная грамматическая информация, необходимая для формирования его буквенного кода при декодировании;
- структура словосочетания приводится к каноническому виду.

Исходными данными для синтаксического анализа служат результаты работы алгоритма морфологического анализа слов. Если слова анализируются с помощью словаря словоформ, то для каждого слова наименования понятия указываются:

- номер канонической формы слова (по словарю словоформ);
- набор переменной грамматической информации (табл. 12), соответствующий данной форме слова;
- постоянная грамматическая информация.

В качестве постоянной грамматической информации для существительных, прилагательных, предлогов, сочинительных союзов и наречий указывается признак принадлежности к соответствующему синтаксическому классу. Кроме того, для существительных указывается признак рода, а для предлогов – перечни падежей, которыми они могут управлять.

Если слова анализируются с помощью словаря основ, то для каждого слова наименования понятия указывается номер канонической формы основы, номер флективного класса и набор переменной грамматической информации. При этом постоянная информация к словам определяется по номерам их флективных классов. Это оказывается возможным благодаря тому, что система классификации слов отражена в нумерации флективных классов (см. табл. 3).

Первым этапом синтаксического анализа словосочетаний является выявление схемы связей между словами, входящими в их состав. Это можно сделать с помощью алгоритма, описанного в разделе 3.5. Но такой способ анализа довольно сложен и обычно применяют более простой способ, основанный на использовании принципа аналогии.

Для синтаксиса *принцип аналогии* формулируется следующим образом: аналогичным последовательностям символов классов слов соответствуют аналогичные схемы синтагматических связей между словами. Под *классом слов* здесь понимается множество слов, обладающих некоторой совокупностью признаков. Для применения этого принципа необходимо выявить все или наиболее часто встречающиеся в текстах последовательности символов классов слов и поставить им в соответствие схемы синтагматических связей. Тогда процесс синтаксического анализа сведется к распознаванию в текстах эталонных последовательностей символов классов и замене их на схемы синтагматических связей. Точность анализа будет зависеть от:

- характера принятой классификации слов;
- длины эталонных последовательностей символов классов слов;

-полноты представления различных синтагматических ситуаций в словаре эталонов.

Она будет тем большей, чем детальнее классификация слов, чем длиннее последовательности символов классов слов в эталонных описаниях синтагматических ситуаций и чем полнее словарь эталонов.

Метод аналогии применяют прежде всего для анализа текстов с ограниченными наборами синтагматических ситуаций, например для анализа именных словосочетаний. С целью оценки эффективности этого метода при синтаксическом анализе именных словосочетаний был обследован словарь научно-технических терминов объемом около 12 000 единиц. При этом выяснилось, что словосочетания, описываемые одинаковыми последовательностями символов обобщенных грамматических классов слов, как правило, имели одинаковые схемы синтаксической связи между словами. Случаи отклонения от этого правила были редкими и составляли менее половины процента.

Вторым этапом синтаксического анализа наименований понятий является определение однозначной грамматической информации к каждому слову. Прежде всего, главному слову словосочетания (первому слева существительному) и определяющим его прилагательным назначается информация «именительный падеж, единственное число», а на прилагательные переносится признак рода главного слова. Далее выполняется операция выделения общей части наборов переменной грамматической информации в группах слов, состоящих из существительного и зависимых от него прилагательных. В результате выполнения этой операции получается либо однозначная грамматическая информация, либо наборы грамматической информации, которые в дальнейшем используются для назначения информации к существительным и прилагательным.

Информация к существительным уточняется в следующем порядке. Если существительное управляется предлогом, то ему назначается первый элемент из соответствующего набора, который содержит информацию о падеже, допустимую для данного типа предлогов (см. табл. 3). Если же существительное управляется другим существительным, то элемент набора выбирается с учетом возможных для такого существительного значений признака падежа. При этом сначала ищется элемент с признаком родительного падежа, затем с признаком творительного и, наконец, с признаком дательного падежа. Информация, выбранная для существительных, распространяется и на подчиненные им прилагательные. Неизменяемым словам словосочетания назначается «нулевая» информация.

Заключительным этапом синтаксического анализа является приведение структуры словосочетания к каноническому виду. При этом выполняются следующие операции:

прилагательные ставятся перед теми существительными, которые они определяют, и упорядочиваются по возрастанию их словарных номеров;

существительные, соединенные сочинительным союзом, располагаются по возрастанию их словарных номеров (в случае необходимости, изменяется расположение слов относительно союза); группы слов, соединенные сочинительным союзом и управляемые существительными, располагаются таким образом, чтобы управляемые слова были упорядочены по возрастанию их номеров; код главного слова словосочетания выносится на первое место.

3.4.2. Кодирование и декодирование наименований понятий

В автоматизированных информационных системах применяются различные способы кодирования понятий.

Под *кодированием понятий* понимают процесс замены их наименований на естественном языке на некоторые формализованные смысловые коды, отражающие содержание этих понятий. Под *декодированием* – обратный процесс перехода от формализованных кодов к наименованиям понятий на естественном языке.

Формализованный код понятия может представлять собой его порядковый номер по заранее составленному инвентарному списку (словарю) или описание его смыслового содержания на некотором формализованном языке. При этом понятие может описываться как объект простой или сложной структуры. Таким образом, и процесс кодирования понятий, и процесс их декодирования являются процессами перекодирования – *перехода от одного способа представления понятий к другому*.

Если понятия кодируются их номерами, то в памяти ЭВМ целесообразно иметь два словаря: словарь слов и словарь пословных кодов словосочетаний (словарь наименований понятий). Первый словарь может быть оформлен в виде словаря словоформ или словаря основ слов. Все его элементы нумеруются.

Во втором словаре каждое наименование понятия представляется сочетанием номеров слов (номер его канонической формы или канонической формы его основы), входящих в его состав, и номером грамматической структуры. *Грамматическая структура словосочетания* содержит информацию о связях между словами и информацию о формах слов, необходимую при декодировании. Различным сочетаниям номеров слов и номеров грамматических структур присваиваются порядковые номера, которые интерпретируются как номера соответствующих понятий.

Автоматическое кодирование понятий осуществляется в три этапа. Сначала отождествляются слова, входящие в наименование понятия, с элементами словаря слов. Слова заменяются их номерами по словарю и сопровождаются грамматической информацией.

На втором этапе кодирования выявляется грамматическая структура наименования понятия (синтаксический анализ). Наконец, полученный в результате первых двух этапов код отождествляется с одним из

элементов словаря наименований понятий и заменяется порядковым номером этого элемента (семантический анализ). Порядковый номер понятия далее используется в качестве его кода.

Отождествление исходных и словарных наименований понятий производится в следующем порядке.

Сначала сочетание номеров слов и грамматическая структура кодируемого наименования понятия ищутся по списку сочетаний номеров слов и по списку грамматических структур словаря понятий и заменяются порядковыми номерами по этим спискам. Далее по номеру понятия из словаря выбирается соответствующий ему номер грамматической структуры и сравнивается с номером, полученным в результате поиска по списку грамматических структур. Если эти номера совпадают, то понятия тождественны друг другу, в противном случае они не тождественны.

Подобно процессу кодирования наименований понятий, их декодирование также осуществляется в три этапа. Сначала по номеру понятия из словаря выбираются соответствующие ему сочетание номеров слов и номер грамматической структуры. Затем из списка грамматических структур извлекается информация о формах слов и их связях, а также корректируется порядок слов в словосочетании (номер главного слова ставится после номеров определяющих его прилагательных). На заключительном этапе формируются буквенные коды словоформ.

Алгоритмы декодирования понятий значительно проще алгоритмов кодирования, особенно если наименования понятий хранятся в основной форме. Если же необходимо согласовать формы наименований понятий с их контекстным окружением, то главному слову и определяющим его прилагательным назначаются соответствующие число и падеж.

Наряду со способами декодирования понятий, основанными на морфологическом синтезе слов, в АИС могут применяться и другие способы.

Можно, например, хранить в памяти машины таблицы соответствия между номерами понятий и их буквенными кодами.

Можно также представить наименования понятий в виде сочетаний номеров словоформ, входящих в их состав, и хранить в памяти машины два словаря – словарь пословных кодов наименований понятий и словарь словоформ. В этом случае декодирование понятий будет производиться в два этапа: сначала, с помощью первого словаря, номера понятий заменяются на их пословные коды, затем, с помощью второго словаря, пословные коды наименований понятий заменяются на их буквенные коды. Последние два способа декодирования понятий очень просты, но их применение связано с необходимостью хранения в памяти машины дополнительных словарей. Кроме того, здесь можно получать только одну форму наименований понятий.

Рассмотренные методы кодирования понятий с автоматическим отождествлением трансформационных вариантов их наименований довольно сложны в реализации и не охватывают всех видов трансформаций. Например, здесь не учитывается возможность

изменения основ слов (например, меры защиты – защитные меры) и возможность изменения схем связей между словами (автоматизированная документальная поисковая система – автоматизированная система поиска документов – система автоматизированного поиска документов). Между тем учет этих явлений весьма желателен, если в АИС не накладывается ограничений на словарь входного языка. Чаще всего это бывает необходимо в документальных системах. Здесь допустимо применение упрощенных способов кодирования, при которых хотя и возможны ошибки, но зато охватывается более широкий класс трансформаций словосочетаний.

3.5. Синтаксический анализ текстов

Рассмотрим принципы автоматического синтаксического анализа текстов на примере алгоритма, который выявляет только поверхностную структуру текстов и является приближенным, но в нем не накладывается никаких ограничений на словарный состав. Анализ текстов здесь проводится по предложениям. Для каждого предложения строится его граф-схема (дерево зависимостей), в которой отображаются буквенные коды слов, связи между словами и грамматическая информация к словам.

В процессе анализа фиксируется лишь факт наличия смысловой связи между словами и направление этой связи (от подчиняющего слова к подчиненному). Более детальная дифференциация связей не производится. Сочинительная связь рассматривается как указание на отсутствие непосредственной связи между словами и словосочетаниями и их подчинение одному и тому же элементу текста.

В качестве материала для анализа использовались научно-технические тексты, на которые не накладывалось никаких ограничений по их словарному составу и синтаксической структуре. Некоторое представление о структуре исходных текстов могут дать следующие их характеристики:

- длина предложений в текстах изменялась в пределах от 6 до 64 слов и составляла в среднем 25 слов;
- более половины предложений были простыми, а сложные предложения включали в свой состав от двух до пяти простых предложений;
- длина интервалов между связанными по смыслу словами (определяемая числом пробелов между ними) колебалась в пределах от единицы до 21 (в среднем она была равна 2,4);
- количество слов, подчиненных одному и тому же слову, изменялось в пределах от нуля до 6 и в среднем составляло 1,55;
- длина цепочек связанных по смыслу слов (измеряемая количеством слов на пути от корня дерева зависимостей к его вершинам) колебалась в пределах от 1 до 10 и в среднем была равна 4.

В рассматриваемой системе синтаксический анализ выполняется за два этапа.

На первом этапе устанавливаются связи между словами внутри небольших фрагментов предложений, границами которых, как правило, являются глаголы, знаки препинания и союзы (исключая знаки препинания и союзы, стоящие между прилагательными, определяющими одно и то же существительное).

На втором этапе устанавливаются связи между упомянутыми выше фрагментами и ищутся «хозяева» для тех слов, для которых они не были найдены на первом этапе.

На первом этапе анализа предложение просматривается с конца с постепенным продвижением к началу. При этом последовательно анализируются пары слов с целью выяснения возможности установления связи между их элементами. Если такая возможность имеется, то связь устанавливают и переходят к следующей паре слов; если нет, то переход к следующей паре осуществляется без фиксации результатов анализа предыдущей пары.

Переход от одной анализируемой пары слов к другой производится по следующим правилам. Если в рассматриваемой паре слов левый элемент является управляющим, то в следующей паре он принимается за правый, а в качестве левого элемента новой пары берется соседнее слово, расположенное слева. Аналогичным образом поступают, когда слова не связаны друг с другом и расположены контактно. Если левый элемент анализируемой пары слов является управляемым, то в качестве левого элемента следующей пары берется слово, расположенное слева от левого элемента анализируемой пары, а правый элемент остается неизменным.

Если элементы анализируемой пары слов не связаны друг с другом и расположены неконтактно, то в качестве правого элемента новой пары слов берется слово, стоящее справа от левого элемента предыдущей пары, а в качестве левого элемента новой пары берется то же слово, которое было левым элементом в предыдущей паре.

Работа второго этапа начинается с членения сложного предложения на простые. Граница между простыми предложениями проводится по знакам препинания и сочинительным союзам. Результаты анализа заносятся в специальную таблицу в памяти ЭВМ, на основании которой могут быть построены гипотезы о составе и структуре текста.

В памяти ЭВМ синтаксические связи между словами оформляются в виде связей между их порядковыми номерами в предложении – каждому номеру слова ставится в соответствие перечень номеров непосредственно подчиненных ему слов. На основе этих исходных данных создается граф-схема предложения.

ЗАДАНИЯ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ

- 1. Составьте логическую схему базы знаний по теме юниты.*

2. Изобразите на схеме основные компоненты АИС.

3. Изобразите на схеме семантическую сеть.

4. Создайте словообразующую парадигму от слова “стучит”.

5. Создайте фрагмент текста, в котором содержится пресуппозиция.

6. Создайте фрагмент текста, в котором содержится явление эллипсиса.

7. Создайте фрагмент текста, в котором содержится инференция.

ЛИНГВИСТИЧЕСКИЕ ОСНОВЫ ИНФОРМАТИКИ

ЮНИТА 1

ЯЗЫК КАК ЗНАКОВАЯ СИСТЕМА

Редактор Н.В. Друх
Оператор компьютерной верстки Д.В. Федотов

Изд. лиц. ЛР № 071765 от 07.12.1998	Сдано в печать
НОУ “Современный Гуманитарный Институт”	
Тираж	Заказ
