

81.2 А+1
П 272

МИНСКИЙ ГОСУДАРСТВЕННЫЙ
ЛИНГВИСТИЧЕСКИЙ УНИВЕРСИТЕТ

На правах рукописи

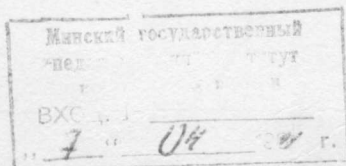
ПЕРЕСЕЛЯК Ирина Васильевна

ИМЕННЫЕ СЛОВСОЧЕТАНИЯ В СТРУКТУРЕ
АНГЛИЙСКОГО НАУЧНОГО ТЕКСТА

Специальность 10.02.19—теория языкознания

Автореферат
диссертации на соискание ученой степени
кандидата филологических наук

Минск 1994



Работа выполнена в Минском государственном лингвистическом университете.

Научный руководитель — кандидат филологических наук, доцент **Т. П. Карпилович.**

Официальные оппоненты — доктор филологических наук, профессор **Ю. Г. Панкрац,**

кандидат филологических наук, доцент **И. Г. Сафьян.**

Ведущая организация — Белорусский государственный университет.

Защита состоится «*26.*» апреля 1994 г. в 14.00 часов на заседании специализированного совета Д. 056.06.01 по присуждению ученой степени доктора филологических наук при Минском государственном лингвистическом университете по адресу: 220662, г. Минск, ул. Захарова, 21.

С диссертацией можно ознакомиться в научном зале Минского государственного лингвистического университета.

Автореферат разослан «*23.*» *марта*. 1994 г.

И. о. ученого секретаря
специализированного совета
доктор филологических наук

З. А. ХАРИТОНЧИК

Ретерируемая диссертация посвящена проблематике оптимизации автоматической обработки научного текста на основании установления функциональной роли именных словосочетаний в его содержательной структуре. В научном тексте, обладающем особой семантикой, именные словосочетания выступают, в основном, как именные терминологические словосочетания (ИТС), семантико-синтаксические свойства которых взаимосвязаны со спецификой отображаемой предметной области (ПО). Анализ информативно-коммуникативной значимости ИТС в тексте может способствовать решению актуальных задач автоматической переработки связных текстов на ЭВМ.

Дискретные языковые элементы текста имеют различное функциональное значение для выражения смысловой информации. В текстах научно-технической литературы особую роль при формировании содержательного аспекта выполняют именные конструкции, т. е. ИТС. Содержательная информация специальных текстов концентрируется в ИТС, служащих для выражения научных понятий, объективирующих систему знаний о некоторой ПО. Изучение ИТС в структуре текстов научных публикаций позволяет установить некоторые важные закономерности функционирования сложных языковых единиц в целостном тексте, от которых зависит решение проблем преобразования текстовой семантики. Извлечение из текста автоматическим путем ИТС, представляющих наиболее информационно-значимые компоненты текста, дает возможность представить его содержание в сжатой форме.

Актуальность данного исследования определяется тем, что проблемы алгоритмического смыслового анализа текстов приобретают большое значение в связи с ростом информационных потоков в современном мире. Необходимость автоматического свертывания смысловой информации научных публикаций предполагает разработку лингвистических моделей извлечения основного содержания научного текста. Установление функциональной роли ИТС при

БИБЛИОТЕКА
МДЛУ

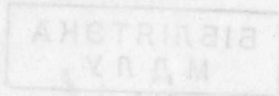
6720

формировании содержательной структуры текста позволяет выделить его наиболее информационно-значимые фрагменты, представляющие в совокупности компрессивное изложение основного содержания целостного текста научной журнальной статьи (НЖС).

Целью работы является изучение информативно-коммуникативного функционирования ИТС в научном тексте и создании на основе лингвистического исследования алгоритма анализа текста НЖС с целью извлечения ее основного содержания. Результаты исследования должны стать основой для разработки действующей экспериментальной вопросно-ответной системы алгоритмического поиска необходимой информации в текстах по специальности "химия протона".

Для достижения поставленной цели в диссертации потребовалось решить следующие задачи:

- построить структурно-семантические модели ИТС рассматриваемой ПО;
- исследовать содержательную структуру текстов НЖС по "химии протона" и проанализировать корреляции между композиционной структурой текстов и их содержанием;
- найти формализуемый способ экспликации семантической структуры текстов и выделить их основное содержание;
- определить функциональную роль ИТС в формировании основного содержания НЖС;
- проанализировать возможности языковой манифестации основного содержания текстов посредством ИТС;
- разработать формализованные методы для автоматического поиска в тексте информационно-значимых ИТС;
- создать лингвистическую информационную базу вопросно-ответной системы, включающую энциклопедические знания о данной ПО;
- провести экспериментальную проверку работы системы на контрольных текстах.



Объектом анализа послужили тексты научных публикаций на английском языке в специальных зарубежных изданиях.

Материалом для исследования являлись 1738 ИТС, выбранных из текстов объемом в 200 тысяч словоупотреблений. Их функционирование было изучено в 20 текстах ПО "химия протона" из английских, американских, канадских и японских журналов по химии.

В ходе исследования были использованы методы семантико-синтаксического и ситуативного моделирования, компонентного анализа, лингвистического описания и алгоритмического моделирования.

Научная новизна данного исследования заключается в том, что впервые структурно-семантические характеристики ИТС рассматриваются в связи с их функционированием в структуре содержания целостного научного текста. Семантическое моделирование текста НХС осуществлено в виде основных аспектов содержания, репрезентированных ИТС. Алгоритмическое обнаружение ИТС, представляющих информативно-коммуникативные элементы содержательной структуры текста, производится комплексом процедур, предусматривающим обращение к энциклопедической информации, т.е. знаниям о ПО "химия протона". Учет позиционных характеристик ИТС в определенных блоках композиционной структуры текста НХС и лексических маркеров, а также анализ синтаксической структуры предложения.

Теоретическая значимость проведенного исследования состоит в разработке проблем семантического анализа целостных текстов научных публикаций и установлении закономерностей функционирования ИТС при формировании содержательной структуры текста.

Практическое значение данной работы заключается в создании действующей экспериментальной системы поиска необходимой информации в текстах НХС по проблемам "химии протона". Система создавалась с учетом прагматических интересов конкретных пользователей, сотрудников Института химии неводных

растворов РАН, и должна способствовать информационному обслуживанию научных исследований. Предложенный подход автоматической переработки текстов может быть использован в более сложных системах информационного поиска, индексирования и реферирования НХС, а также при компьютерном обучении студентов реферированию иноязычного текста.

На защиту выносятся следующие основные положения диссертации:

1. Тексты НХС обладают определенным языковым построением, репрезентирующим некоторый фрагмент научного знания, где главным средством передачи семантической информации служат наименования понятий, т.е. ИТС, выступающие в качестве основного средства номинации единиц знания.

2. Представляя элементы системного знания о некоторой ПО, ИТС в тексте взаимосвязаны логическими отношениями, определяющими способ семантического структурирования текстов.

3. Экспликация структуры содержания текстов может быть проведена посредством выявления основных аспектов содержания, которые представляют собой инвариантную модель для семантического анализа тематически однородных текстов.

4. Обнаружение в тексте ИТС, передающих основные аспекты содержания, позволяет выделить информационно-значимые фрагменты текстов, в совокупности представляющих основное содержание научных публикаций.

5. Разработка формальных процедур отождествления в тексте ИТС, выражающих основные аспекты содержания, основывается на ряде методов, включающих в себя обращение к энциклопедическим знаниям, позиционные характеристики фрагментов текста, учет лексических маркеров, анализ синтаксической структуры предложения.

А п р о б а ц и я р а б о т ы. Основные положения и резу-

льтаты исследования были представлены на научных конференциях Минского ГПИЯ, Ивановской ГХТА, ИвГУ, ИвТИ и в публикациях автора.

Цель и задачи диссертации определили ее структуру: диссертация состоит из введения, трех глав и заключения, списка использованной литературы и словарей, а также приложений.

Во введении обосновывается актуальность темы, определяются цель, задачи и методы исследования, отмечается новизна, теоретическая и практическая значимость, излагаются основные положения, выносимые на защиту.

В первой главе рассматривается проблема определения текста в лингвистике, особенности научно-технических текстов, знаковость текста и его семантика. Предлагается способ экспликации семантической структуры текстов в виде перечня содержательных аспектов, воспроизводящих основное содержание текстов. Изучается проблематика ИТС и их роль в формировании содержания текстов.

Вторая глава посвящена изучению структурирования содержания научных статей по проблемам "кинний протона", а также синтаксическим и семантическим особенностям ИТС данной ПО и их участию в передаче основных аспектов содержания текстов.

В третьей главе проводится анализ композиционной структуры текстов НТС в сопоставлении с содержащейся в ней семантической информацией. Выделяются способы формального определения в тексте ИТС, передающих основное содержание текстов, описывается общий алгоритм работы системы и ее лингвистическая информационная база, а также ряд отдельных алгоритмов для анализа текстов.

В заключении приводятся выводы обобщающего характера и результаты исследования, намечаются возможные перспективы дальнейшей работы по данной проблематике.

Содержание работы

Определение онтологических признаков текста предполагает изучение коммуникативно-функциональной стороны отдельных текстовых компонентов. Элементы языковой системы в структуре текста подчиняются задаче выражения законченного содержания, отвечающего, целям коммуникации. Обладая многоплановым характером, текст репрезентирует результаты речемыслительной деятельности в языковой форме, т.е. представляет некое знание о действительности в комбинациях языковых знаков.

Содержание текста эксплицируется в языковых средствах, имеющих дискретный и линейный характер, которые не находятся во взаимно-однозначных соответствиях с семантическими явлениями. Семантика текста представляет собой сложный информативный комплекс, интегрирующий замысел автора текста в соответствии с условиями и целями коммуникации. Объективируя некоторый фрагмент знаний, содержательный аспект текста характеризуется целостностью и структурностью, так как смысловая структура текста складывается, с одной стороны, из отношений объектов денотативной сферы, а с другой стороны, она отражает логические взаимосвязи между коммуникативными действиями.

Семантика научного текста, как и любого другого, многомерна и неоднородна, поскольку включает в себе не только предметно-тематический компонент, но и отношение субъекта познания к этому предметному содержанию, а также отношение самого текста к описываемой реальности. Преобладание в научном тексте когнитивного (концептуального) компонента определяет метаязыковую форму интегрированных ментальных и коммуникативных процессов. Пересечение когнитивных, коммуникативных, ситуативно-прагматических компонентов обуславливает сложную семантическую структуру научного текста, обнаружение которой необходимо для моделирования процесса извлечения информативно-значимых элементов текста.

Информационное обслуживание научных исследований предусматривает семантическую переработку связанных текстов научных публикаций в аннотации и рефераты, позволяющие в краткой форме представить основное содержание целостных текстов. Особенно актуальной считается переработка иноязычной информации, помогающая преодолеть языковые барьеры для специалистов, не владеющих иностранным языком. В этом случае проблема сжатия текста дополняется задачей осуществления перевода с языка оригинала на русский язык.

Автоматическое реферирование (АР), основанное на использовании смысловых преобразований — перспективное направление теоретических и прикладных исследований, требующее разнообразных подходов и экспериментальной проверки. В подобных системах прежде всего решается задача экспликации структуры содержания первичного текста формальными способами (И. П. Севбо, Э. Ф. Скороходько, Р. Г. Пиотровский, А. В. Зубов, Л. Н. Беязева, А. И. Новиков, Х. А. Арзикулов, В. Е. Берзон, С. Н. Гиндин, Н. Н. Леонтьева, Б. В. Якушин). На основании установления неравнозначности элементов содержания текста можно выделить его значимые компоненты, т. е. получить модель основного содержания текста.

Способы выявления структуры содержания текстов различны в зависимости от реализуемых лингвистических подходов. Одним из подходов к экспликации семантической структуры текстов и обнаружения стабильных элементов предметных ситуаций и отношений между ними можно считать выделение аспектов содержания определенного класса текстов. Его преимуществом считается то, что анализ текстов осуществляется исходя из содержания конкретных текстов по той или иной проблематике. Объективирование основного содержания текстов происходит на основании конкретных задач автоматической переработки текстов и целей потребителей информации. Выделение инвариантных элементов содержания текстов при помощи аспектов содержания является легко формализуемым методом для

экспликации содержательной структуры текстов, позволяющим представить ее в виде перечня основных элементов содержания для некоего множества тематически однородных текстов.

Научный текст, наиболее распространенным представителем которого является НХС, обнаруживает системно-структурные признаки, которые проявляются в иерархичности его семантической структуры, находящей выражение во внешней организации текста, хотя отношения, связывающие элементы структуры содержания могут иметь и имплицитный характер.

Языковая форма английских научных текстов характеризуется наличием большого количества номинативных конструкций, абстрактной лексики, особых способов словообразования, усложненных синтаксических структур словосочетаний и предложений. Строгая композиционная структура, закономерность построения сверхфразовых единств и абзацев, а также преимущественное употребление научных терминов, значение которых определено заранее в логико-понятийной системе знания определенной ПО, детерминируют тенденции к стандартизации текстовых форм, к формализации средств научной коммуникации, позволяющие повысить степень информативности и однозначности научного текста. Эти свойства научного текста являются важной предпосылкой его аналитико-синтетической переработки в системах человеко-машинной коммуникации.

В процессе реферирования, основанном на выделении ИТС, необходимо определить в тексте словосочетания с наибольшей семантической нагрузкой, способные замещать значительные фрагменты содержания. Наиболее емкими оказываются ИТС, обозначающие модели микроситуаций в свернутом виде, так как ИТС обладают способностью к латентной предикации. Образование этих синтаксических единиц обусловлено объективными отношениями предметов и явлений реальной действительности. Познание действительности начинается с выделения материальных субстанций и определения их признаков. Номинация

новых структур познания реализуется вследствие комбинаторных возможностей языковых знаков, что приводит к созданию именных словосочетаний различной устойчивости.

В предложении и тексте устойчивые именные словосочетания приобретают коммуникативную достаточность, они обозначают предметы и процессы в их действительных связях, соотносясь с реальной ситуацией и прагматическим фокусом сообщения. Именно в тексте именные словосочетания обнаруживают системные взаимосвязи, обуславливаемые логикой воспроизводимой действительности.

В английском языке структура именных словосочетаний проявляется в форме именных комплексов, где синтаксические зависимости не выражены морфологически. Они заключают в себе высокий семантический потенциал, обладая способностью передавать разнообразные смысловые отношения имени и альянкта.

Выступая как тематические доминанты научного текста, ИТС функционируют как структурно-семантические единицы содержательной структуры текста. Присутствующая в тексте имплицитно иерархия системных отношений между терминами определяет структуру темо-рема-тической прогрессии в тексте, где каждое ИТС актуализируется в зависимости от его информационной значимости. Функционирование ИТС в тексте детерминировано не только их системными свойствами, закрепленными в логико-понятийной системе знаний ПО, но и авторским замыслом, коммуникативными целями и требованиями к языковой организации конкретного вида текста. Проблема соотношения содержательной структуры текста с его языковыми конститuentами может быть решена как выделение ИТС, представляющих основные аспекты содержания текстов. Этот процесс требует моделирования системы знания некоторой ПО, а также тех проблемных ситуаций, которые воспроизводятся в текстах определенного типа. Аспекты содержания текстов, представленные в виде ИТС, которые передают основное содержание текстов, воспроизводят устойчивую модель семантической структуры

текстов, объективируя логические связи ИТС в научном тексте.

Изучение семантической структуры текстов ПО "химия протона" позволило выделить ряд аспектов содержания, являющихся аналогом основного содержания этих текстов. К ним относятся: "тема и предмет исследования", "анализ предыдущих исследований", "цель исследования", "типы и виды химических реакций", "участники реакции (кислоты и основания)", "растворитель и его свойства", "вещества для синтеза", "препаративные методики", "теоретические и экспериментальные методы изучения реакций", и т. д.

Выделенные аспекты содержания НЖС по "химии протона" можно разделить на две группы. К первой относятся те аспекты, которые связаны с презентацией исследовательской деятельности ученых, когда устанавливается предмет изучения, его связи с ранее проведенными исследованиями, цель нового исследования и анализируются полученные результаты. Эти элементы содержания текста обладают особым текстообразующим статусом и выражены в текстовых фрагментах, которые представляют собой, как правило, свободные комбинации слов и ИТС в составе целых предложений с разнообразными полипредикативными структурами.

Вторую группу образуют аспекты, описывающие знания об объектах, явлениях и процессах ПО "химия протона". Эти аспекты отражают рубрики тезауруса данной области знания и моделируют логико-понятийную систему данной ПО. Они манифестированы в тексте как в виде устойчивых наименований понятий, т. е. терминов или ИТС, так и в виде комбинаций терминов, т. е. свободных словосочетаний.

Явление неоднородности содержательных аспектов научного текста, проявляющееся в их различной текстовой манифестации, отображает дифференциацию субъектно-объектных отношений в процессе познания реального мира. Эти отношения воспроизводят структуру взаимосвязей как между познающим субъектом и объектом поз-

нения, так и между разными объектами описываемого фрагмента действительности.

Полученные в результате анализа текстов ИТС были распределены с помощью экспертов на группы в соответствии с их принадлежностью к различным аспектам предметного содержания текстов. Данные структурного анализа ИТС приведены в тексте диссертации в нескольких таблицах (например, табл. 1). Классификация по количеству составляющих ИТС компонентов позволила выделить наиболее частотные синтаксические модели для всех аспектов содержания. Наиболее распространенной моделью словосочетания в английских текстах по "химии протона" является субстантивное словосочетание $N1+N2$ (46%) -vibration frequency, distribution curve, что подтверждает данные об универсальности этой модели в английском терминообразовании. Второй наиболее употребительной моделью двухкомпонентных ИТС в изучаемом подязыке является модель $A+N$ (19%), служащая для экспликации предметно-признаковых отношений: stoichiometric concentration, molar absorptivity.

Среди трехкомпонентных словосочетаний (24%) наиболее употребительны модели $N1+N2+N3$, $A+N1+N2$ и $A1+A2+N$, например: solvent exchange equilibrium -равновесие обмена растворителя, high polarity solvents -высокополярные растворители, rigid chemical bonds -жесткие химические связи. Наибольшие комбинационные потенции проявляют именно трехкомпонентные модели, очевидно, это связано с оптимизацией процессов номинации, причем в них часто фигурируют причастия и наречия, например: differently solvated anion -различно сольватированный анион.

Синтаксические структуры ИТС в различных аспектах содержания носят универсальный характер, что свидетельствует о недостаточности структурных характеристик ИТС для алгоритмического выделения словосочетаний, описывающих различные аспекты содержания изученных текстов. Однако этот этап работы позволял систематизировать

Таблица 1 (фрагмент)

Структурные модели ИТС, описывающих содержательный аспект "Параметры реакции"

Структура с/с	Ком-во	Пример	Перевод
Двухкомпонентные			
1. $N_1 + N_2$	208	transfer enthalpy activation parameters first-order conditions	энтальпия переноса активационные параметры условия первого порядка
2. $A + N$	72	semiquantitative correlation	полуколичественная корреляция
3. $N_1 + p \cdot N_2$	31	energy of activation rates of detritiation	энергия активации скорость детритирования
Трехкомпонентные			
1. $A + N_1 + N_2$	35	intermolecular separation coordinate	межмолекулярная координата разделения
2. $N_1 + N_2 + N_3$	29	transition state structure ion association constants	структура переходного состояния константы ассоциации ионов
3. $N_1 + p \cdot N_2 + N_3$	12	rates of deuterium transfer entropy of activation effect	скорость переноса дейтерона эффект энтропии активации
4. $A + N_1 + p \cdot N_2$	4	structural change of activation	структурное изменение активации
5. $N_1 + p \cdot A + N_2$	3	concentration of free water	концентрация свободной воды
6. $A_1 + A_2 + N$	2	combined kinetic shift	суммарная кинетическая энергия
7. $N_1 + A + N_2$	2	п.м.т. chemical shift	химический сдвиг а.м.т.
8. $N_1 + N_2 + p \cdot N_3$	1	excess heat of mixing	избыточная теплота смешивания
Всего:	414		

корпус изученных ИТС и, в дальнейшем, заложить их в лингвистическую информационную базу компьютерной системы.

Следующим этапом изучения ИТС, относящихся к различным аспектам содержания, было исследование семантики ядерного существительного ИТС, которое могло дать семантические критерии для определения принадлежности того или иного ИТС к какому-либо содержательному аспекту. Исследование семантики ядерных компонентов ИТС, являющееся одним из апробированных методов прикладной и теоретической лингвистики, проводилось на основании выделения универсальных логико-семантических категорий, анализа данных толковых и энциклопедических словарей и изучения микроконтекста словосочетаний.

Были получены семантические признаки (Сем П), характеризующие значения ядерного существительного ИТС, например, Сем П- вещество (acid, base, alcohol, amine), элементарная частица (proton, neutron, electron), физико-химический процесс (abstraction, activation, absorption, vibration), структурная единица вещества (atom, ion, group, molecule), физическая величина (mass, viscosity, energy, potential), и т. д. Всего было выделено 33 Сем П для изученных ИТС.

Сравнение и сопоставление семантических признаков ядерного компонента и определений ИТС, входящих в различные содержательные аспекты, показало, что полученных Сем П достаточно для однозначного определения лишь немногих аспектов. Большая же часть изученных аспектов проявляет явления пересечения и совпадения Сем П. Это касается таких аспектов содержания, которые относятся к теоретическим сторонам ПО "химия протона", что можно объяснить взаимосвязями процессов и явлений, описываемых в различных содержательных аспектах, т. е. некоторой диффузностью системы концептуальных представлений сложной предметной области. Языковые средства для описания подобных систем оказываются полифункциональными и лабильными, а их семантическая интерпретация является относительной

вследствие постоянного изменения научного знания о реальной действительности. Многозначность языковых единиц, употребляемых для обозначения научных понятий, снимается за счет ограничений, накладываемых рамками контекста и процессами актуализации значений словосочетаний в тексте.

Таким образом, структурные и семантические характеристики ИТС оказываются недостаточными для идентификации их роли в структуре содержания текста, т.е. принадлежности к тому или иному аспекту содержания, поэтому, как представляется, только изучение их функционирования в тексте может привести к отождествлению необходимой информации.

В тексте НЖС наблюдаются определенные закономерности распределения содержательной информации в блоках композиционной структуры, что позволяет установить пути оптимального поиска информационно-значимых ИТС, передающих в тексте единицы знания. Композиция НЖС выступает как совокупность структурно-семантических признаков, определяющих способы кодирования и декодирования научной информации. Сегментирование текста на отдельные структурно-семантические блоки возможно благодаря их автосемантической и относительной завершенности в пределах целого текста.

Структурные элементы научной статьи выполняют определенную коммуникативно-прагматическую функцию. Заголовок статьи, аннотация, введение, основная часть и заключение реализуют коммуникативное намерение автора и адекватно воспроизводят содержательные элементы научной информации. Данные о распределении содержательной информации текста НЖС в отдельных композиционных блоках представлены в табл. 2. Из табл. 2 следует, что различные аспекты содержания НЖС имеют неодинаковые позиционные характеристики. Аспекты, относящиеся к презентации структуры исследовательской деятельности, обнаруживают тенденцию к четкой локализации в определенных блоках композиционной структуры НЖС, например, в заголов-

ке и во введении, поэтому их алгоритмический поиск можно осуществить на основе позиционных характеристик.

Таблица 2 (фрагмент)

Корреляция аспектов содержания статей и их позиций
в композиционной структуре текста

Разделы статьи		Заголовок	Аннотация (Abstract)	Введение (Introduction)	Эксперимент (Experimental)	Результаты (Results)	Обсуждение (Discussion)	Выводы (Conclusion)	Библиография
субъектная деятельность	1. Тема и предмет исследования	+	+	+					
	2. Анализ предыду- щих исследований			+					+
	3. Цель исследо- вания			+					
	4. Выводы						+	+	
	5. Типы и виды реакций	+	+	+		+	+	+	
	6. Кислоты и их свойства	+	+	+	+	+	+	+	
	7. Основания и их свойства	+	+	+	+	+	+	+	
	8. Растворитель и его свойства	+	+	+	+	+	+	+	
	9. Продукты реакции		+	+		+	+		
объект исследования	1. Тема и предмет исследования								
	2. Анализ предыду- щих исследований								
	3. Цель исследо- вания								
	4. Выводы								
	5. Типы и виды реакций								
	6. Кислоты и их свойства								
	7. Основания и их свойства								
	8. Растворитель и его свойства								
	9. Продукты реакции								

Однако аспекты содержания, передающие концептуальное описа-

ние ПО "химия протона", т.е. объект исследования, находятся во всех композиционных блоках НЖС и их однозначное отождествление в тексте невозможно сделать на основании только лишь позиционных критериев. Композиционная структура НЖС позволяет сузить сферу алгоритмического поиска необходимой информации, но для большинства аспектов содержания необходимы дополнительные критерии.

Особую роль выполняют ИТС, находящиеся в сильных позициях текста, например, в заголовке, в первом предложении аннотации, в первом предложении текста, в конце абзацев, разделов и всего текста. ИТС в заголовке текста раскрывают тему и предмет исследования. ИТС в первом предложении аннотации манифестируют тематическую доминанту последующего текста, например, в тексте с заголовком "Kinetic isotope effects in the reaction of 4-nitrophenylnitromethane with various bases in chlorobenzene" тема исследования обозначена в виде наименования изучаемой реакции - reaction of 4-nitrophenylnitromethane with various bases in chlorobenzene - с конкретным указанием всех участников реакции. Первое предложение аннотации к этому тексту: "The kinetics of the proton-transfer reactions of 4-nitrophenylnitromethane with various bases in chlorobenzene solution have been studied, and deuterium isotope effects determined" содержит обозначение предмета и темы исследования посредством повторения номинативных элементов заголовка с дополнениями и уточнениями: proton-transfer reactions называет тип реакции, 4-nitrophenylnitromethane - кислоту, various bases - основания, chlorobenzene solution - растворитель, deuterium isotope effects - обозначает предмет исследования в данном тексте.

В первом предложении основного текста, обладающем автосемантическими свойствами, не только повторяется тема и предмет исследования, но и устанавливается логическая связь предыдущих

исследований с данным конкретным изучением проблемы. излагаемой в тексте. Например: "Previous investigations of the proton-transfer reaction of 4-NPPCM and substituted derivatives with TMG studied the activation parameters, deuterium isotope effects, and the Hammett correlation".

Для формализованного обнаружения необходимых фрагментов текста используются также определенные лексические показатели, называемые лексическими маркерами. Они позволяют обнаружить необходимую информацию на уровне отдельного предложения. Входя в его состав, лексические маркеры образуют вспомогательную подсистему текста, обрамляющую собственно содержательную информацию. Изучение текстов НЖС по "химии протона" подтвердило, что с помощью лексических маркеров можно довольно точно найти предложения, описывающие те аспекты содержания, где представлена структура субъектной деятельности, т.е. "анализ предыдущих исследований", "цель исследования" и "выводы". например, маркеры this work continues, we now present, the present work, the present investigation, будучи дополненными позиционными критериями, указывают на аспект "цель исследования".

В некоторых случаях для поиска необходимых фрагментов текста, относящихся к предметному содержанию, проводится анализ синтаксической структуры предложения. Он основывается на темо-рематических отношениях в структуре предложения, когда ИТС, несущие необходимую информацию, отождествляются на основании выделения формальных границ сказуемого или же других признаков.

Привлечение энциклопедической информации (табл. 3) необходимо для всех аспектов, относящихся к описанию ПО "химия протона". Представление знаний, увеличивающее поисковые возможности системы, может осуществляться различными способами, что определяется как свойством самой ПО, так и задачами информационного поиска, а также требованиями к размещению информации оптимальным способом в

Способы идентификации информационно-значимых
фрагментов текста

Способы Идентификации	Позици- онные	Лексичес- кие марке- ры клише	База энци- клопедичес- кой информ.	Анализ синта- ксич. СТРУКТ.
Аспекты содержания				
1. Тема и предмет исследования	+			
2. Анализ предыдущих исследований	+	+		
3. Цель исследования	+	+		
4. Выводы	+	+		
5. Типы и виды реакций	+			
6. Кислоты и их свойства	+		+	
7. Основания и их свойства	+		+	
8. Растворитель и его свойства	+		+	
9. Продукты реакции	+		+	+
10. Вещества для синтеза	+		+	+
11. Препаративные методи- ки	+		+	
12. Экспериментальная аппаратура	+		+	
13. Теоретические и эк- спериментальные мето- ды изучения реакции	+		+	
14. Изотопный эффект	+		+	
15. Параметры реакции	+		+	
16. Механизм реакции	+		+	

памяти машины. В нашем случае была выбрана модель Фрейма, поскольку в центре внимания специалистов по "химии протона" находится стандартный процесс-химическая реакция, все составляющие которой можно зафиксировать в качестве узлов Фрейма. При помощи Фрейма задается модель ситуации, являющаяся исходной для формирования текстовых ситуаций и характеризующаяся категориальной устойчивостью ПО. Узлы Фрейма соответствуют тем аспектам содержания, которые представляют логическую структуру знания об объектах и процессах рассматриваемой ПО. Фактически они соответствуют рубрикам тезауруса данной ПО и открывают доступ к спискам ИТС, являющихся наименованиями базовых понятий этой специальности. Поиск и отождествление этих понятий в тексте приводит к обнаружению ИТС, являющихся ответами на запросы пользователей (рис.1).

Алгоритмический поиск (рис.2) в тексте ответов на запросы пользователей, которые соответствуют выделенным аспектам содержания, осуществляется по нисходящему принципу-от анализа структуры целого текста до отдельных его фрагментов-комплексом поисковых процедур, которые включают в себя позиционные критерии, лексические маркеры, обращение к энциклопедической информации и анализ синтаксической структуры предложения. Выделяемые в тексте ИТС репрезентируют информационно-значимые фрагменты текста, в совокупности передающие его основное содержание.

Система автоматического поиска необходимой информации в тексте базируется на лингвистических данных, заложенных в лингвистическую информационную базу, функционирующую в виде автоматического словаря с определенной структурой. Для вопросно-ответной системы поиска информации в тексте был создан автоматический словарь ИТС, так как в данной системе именно ИТС являются основной единицей языкового анализа текстов и выдаются пользователю в качестве ответов на его запросы. Автоматический словарь словосочетаний отвечает задачам системы и позволяет за-

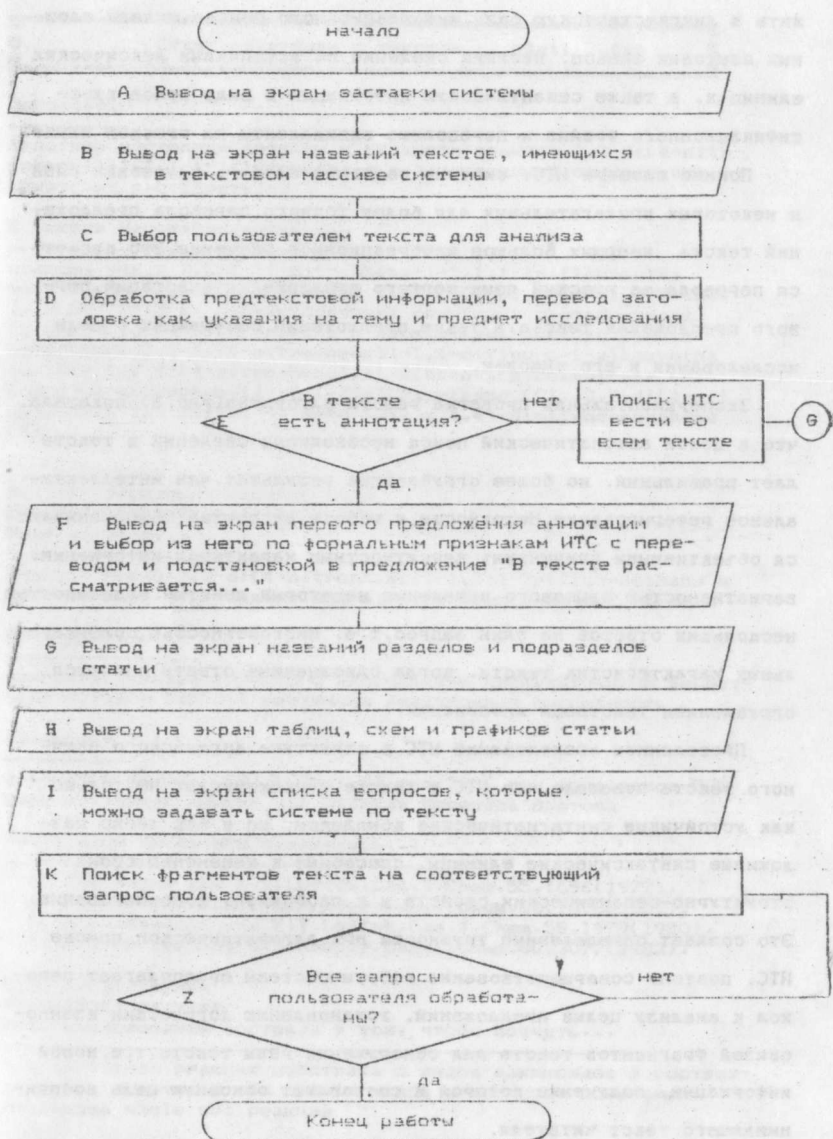


Рис.2. Общий алгоритм работы пользователя с диалоговой системой "ПРОТОН"

жить в лингвистическую базу информационного поиска модели сложных языковых знаков, несущих сведения об устойчивых лексических единицах, а также семантическую информацию в виде кодов классификационного Фрейна и переводные эквиваленты на русском языке.

Помимо словаря ИТС, система содержит словарь служебных слов и некоторых прилагательных для более точного перевода предложения текста, имеющих большое информационное значение. Это касается перевода на русский язык первого предложения аннотации, первого предложения текста, а также предложений, сообщающих о цели исследования и его выводах.

Экспериментальная проверка работы алгоритма (рис. 3) показала, что в целом автоматический поиск необходимых сведений в тексте дает правильный, но более огрубленный результат, чем интеллектуальное реферирование. Неточности в работе алгоритма обуславливаются объективными причинами: вероятностным характером информации, вариативностью языкового выражения некоторых понятий, возможностью нескольких ответов на один запрос, т.е. многомерностью содержательных характеристик текста, когда однозначные ответы являются огрублением текстовой информации.

Проведенное исследование ИТС в структуре английского научного текста показало, что ИТС в тексте функционируют не только как устойчивые синтагматические комплексы, но и как легко разложимые синтаксические единицы, способные к изменению своих структурно-семантических свойств и к свободному комбинированию. Это создает определенные трудности при алгоритмическом поиске ИТС, поэтому совершенствование работы системы предполагает переход к анализу целых предложений, установлению логических взаимосвязей фрагментов текста для обнаружения ремы текста, т.е. новой информации, получение которой и составляет основную цель воспринимающего текст читателя.

Основные положения диссертации отражены в следующих публи-

Отчет по работе с текстом - _stat1 .txt

Заголовок

кинетика изотопные эффекты и механизм реакции 2,2-di(4-nitrophenyl)-1,1,1-trifluoroethane с алкоксидными основаниями в спиртовых растворителях

В тексте рассматривается...

реакция между 2,2-di(4-nitrophenyl)-1,1,1-trifluoroethane и алкоксидными основаниями OCN_3 , OC_2H_5 , ONC_4H_9 , $\text{OCH}(\text{CH}_3)_2$ и $\text{OC}(\text{CH}_3)_3$ в соответствующих спиртовых растворителях является многоступенчатой реакцией с несколькими промежуточными веществами 2,2-di(4-nitrophenyl)-1,1-difluoro-1-alkoxyethane (A), 2,2-di(4-nitrophenyl)-1-fluoro-1-alkoxyethane (B), 2,2-di(4-nitrophenyl)-1,1-dialkoxyethene (C), 2,2-di(4-nitrophenyl)-1,1-difluoroethene (D), и 4,4'-dinitrobenzophenone (E)

Анализ предыдущих исследований

Были изучены...

реакция между 2,2-di(4-nitrophenyl)-1,1,1-trifluoroethane и алкоксидными основаниями has already been examined /1-4/ и

Были найдены...

многоступенчатый процесс с несколькими промежуточными веществами который exhibit различная реакционная способность

Однако...

изотопный эффект $k(\text{H})/k(\text{D})=1,42$ при 26°C при appearance ^52S peak was очень низкий для реакции переноса протона

Ранее этой проблемой занимались...

1. J.Kurzawa and K.T.Leffek.Can.J.Chem.55,1696(1977).
2. K.T.Leffek and G.Schrader.Can.J.Chem.60,1696(1982).
3. A.Jarczewski and K.T.Leffek.Can.J.Chem.58,1979(1980).
4. K.T.Leffek and G.Schrader.Can.J.Chem.60,3077(1982).

Цель исследования

Цель исследования состояла в том, чтобы изучить...

was initiated реакция субстрата с рядом алкоксидов в соответствующих спиртах чтобы obtain ясное и логичное понимание механизма whole set реакций

Препаративные методики

Aldrich, bp 165-165°C

Рис.3.фрагмент результата работы диалоговой системы "Протон"

кациях автора:

1. Роль частотного терминологического словаря в оптимизации обучения лексике // Проблемы методики преподавания иностранных языков в неязыковом вузе. Сб. науч. ст. - Иваново: ИХТИ, 1987. - С. 255-262. Деп. в НИИВШ. N 115-88 от 07.01.1988.
2. Семантическое моделирование терминологических словосочетаний // Лексикографические методы в обучении и научных исследованиях. Тезисы межвуз. конф. - Иваново: ИХТИ, 1988. - С. 33-34.
3. Проблемы применения ЭВМ в обучении иностранному языку в неязыковом вузе // Применение технических средств в преподавании иностранных языков. Тезисы межвуз. конф. - Новочеркасск: НПИ, 1989. - С. 47.
4. Семантика многокомпонентных субстантивных словосочетаний терминологического характера // Лексикографические методы в обучении и научных исследованиях. Сб. науч. ст. - Иваново: ИХТИ, 1990. - Деп. в ИНИОН АН СССР. - С. 177-183. - N 41730 от 28.04.1990.
5. Терминологические именные словосочетания в целостном тексте научной публикации // Совершенствование преподавания иностранных языков в высшей и средней школе. Тезисы респуб. конф. - Душанбе: ДГПИ им. Т. Г. Шевченко, 1991. - С. 38-39.
6. Именные словосочетания в структуре научного текста // Белорусский язык среди европейских. Тезисы докладов аспирантов и студентов. - Мн.: МГПИИЯ, 1992. - С. 46.
7. Семантические особенности определяемого существительного в именных словосочетаниях терминологического характера // Проблемы словосочетания. Тезисы межвуз. конф. - Пятигорск: ПГПИИЯ, 1992. - С. 183-184.
8. Корреляция композиционного членения научного текста и информационно-значимых аспектов его содержания // Белорусский язык среди европейских. Тезисы докл. аспирантов и студентов. - Мн.: МГЛУ, 1993. - С. 75.

Перес