Московский государственный
университет им. М.В. Ломоносова
Факультет вычислительной математики и кибернетики

Волкова И.А.

Введение в компьютерную лингвистику. Практические аспекты создания лингвистических процессоров

(Учебное пособие для студентов факультета ВМиК МГУ)

Москва
2006

УДК 519.6+681.3.06

Данное учебное пособие разработано в поддержку спецкурса «Компьютерная лингвистика», читаемого на факультете ВМиК для студентов 3-5 курсов. Приводятся подробные пояснения и рекомендации.

Рецензенты:

проф. Машечкин И.В.

доц. Корухова Л.С.

Волкова И.А.

«Введение в компьютерную лингвистику. Практические аспекты создания лингвистических процессоров. (Учебное пособие для студентов факультета ВМиК МГУ)»

Издательский отдел факультета ВМиК МГУ
(лицензия ЛР №040777 от 23.07.96), 2006 — 43 с.

Печатается по решению Редакционно-Издательского Совета факультета Вычислительной Математики и Кибернетики МГУ им. М.В. Ломоносова.

ISBN 5-89407-242-5

Содержание

1. Основные понятия и определения компьютерной лингвистики.

Компьютерная лингвистика изучает различные аспекты (теоретические, алгоритмические, программистские), связанные с реализацией всевозможных систем, обрабатывающих какие либо высказывания на ЕЯ (ЕЯ-систем).

Можно выделить следующие основные классы ЕЯ-систем.

Интеллектуальные вопрос-ответные системы. При разработке этих систем основное внимание уделяется развитию моделей и методов, позволяющих осуществлять перевод высказываний на ЕЯ, относящихся к узким и заранее фиксированным проблемным областям, в формальное представление, интерпретацию этих высказываний и генерацию ответных высказываний на ЕЯ по заранее известным, фиксированным правилам.

Системы общения с базами данных. Основная задача таких систем заключается в выполнении перевода запросов неподготовленных конечных пользователей базы данных с ЕЯ на формальный язык запросов к базе данных.

Диалоговые системы решения задач. Эти системы берут на себя не только функции доступа к базе данных, но и функции интеллектуального монитора, обеспечивающего решение заранее определенных классов задач (например, планирование путешествий, составление контрактов). Основное направление использования этих ЕЯ-систем — реализация естественноязыкового общения с экспертными системами.

Системы обработки связных текстов. Эти системы занимаются обработкой текстовой информации и речи. Объем и разнообразие такой информации возрастает с каждым днем. Развитие и совершенствование систем автоматической обработки текстов на ЕЯ (АОТ-ситем) в настоящее время является наиболее актуальным и перспективным. Примеры областей применения АОТ-систем: обучение естественному языку, автоматический перевод, автокорректоры, распознавание речи, синтез речи, автоматическое реферирование, поисковые системы.

Практически любые ЕЯ-системы в той или иной форме имеют и используют морфологический компонент ЛП, некоторые из них так или иначе используют и синтаксический компонент ЛП. Наиболее развитые и сложные ЕЯ-системы имеют в своем составе также семантический и прагматический компоненты и анализируют не только отдельные предложения, но и входной текст в целом.

Лингвистический процессор (ЛП) — комплекс программ, обеспечивающий анализ и синтез текстов на естественном языке. Задачей ЛП является разбор и «понимание» поступившей на вход фразы на ЕЯ (в случае анализа) или построение фразы ЕЯ, соответствующей формальному описанию ее смысла (в случае синтеза).

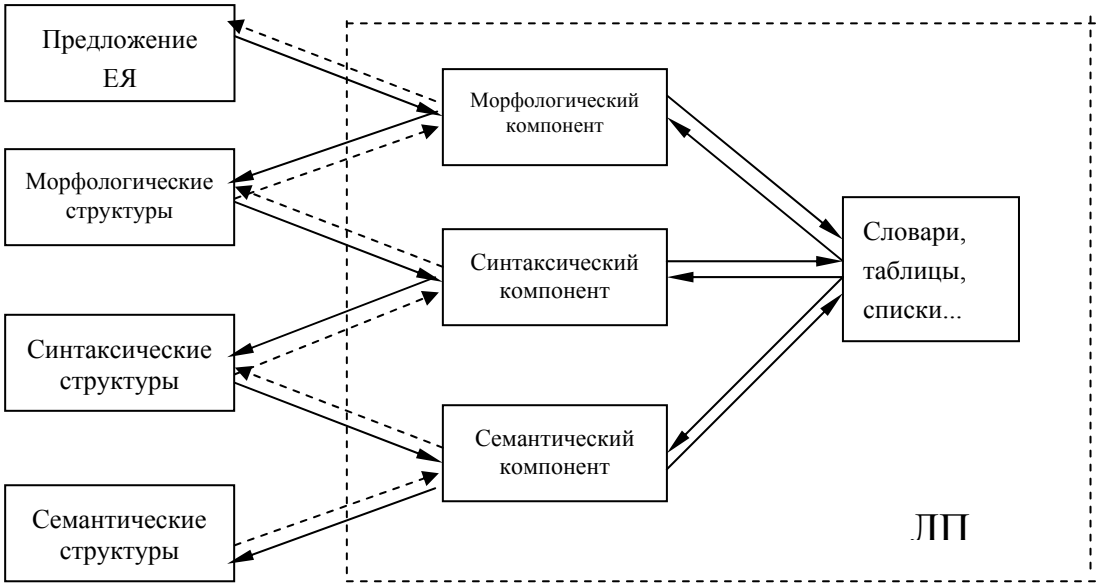
В состав ЛП входят три вида обеспечения:

- лингвистическое (формальная модель ЕЯ, словари, грамматики, лингвистические таблицы, правила);
- математико-алгоритмическое (трансляторы формальных языков, алгоритмы переработки текстов):
- программное.

Восприятие естественноязыковой информации машиной в широком смысле заключается в распознавании смысла текста, которое осуществляется на основе автоматических словарей и формальных грамматик.

Текст можно определить как избыточную многоуровневую систему хранения и передачи информации.

Традиционно в ЕЯ выделяются следующие языковые уровни: морфологический, синтаксический, семантический (иногда и прагматический). На каждом языковом уровне используются свои структуры данных, которые обрабатываются и формируются соответствующими компонентами ЛП. В целом ЛП можно рассматривать как многоуровневый транслятор ЕЯ, переводящий (в случае анализа) входное предложение ЕЯ во внутреннее представление смысла этого предложения и наоборот в случае синтеза. ЛП работает по следующей схеме:



Существует два аспекта системного изучения языка, противопоставляющиеся по типу изучаемых отношений между единицами языка и/или языковыми конструкциями: парадигматика и синтагматика.

Парадигматика — раздел науки о языке, занимающийся парадигматическими отношениями (их классификацией, определением области действия и т.п.). Другими словами, парадигматика изучает языковую систему как совокупность лингвистических классов — парадигм.

Парадигма — любой класс лингвистических единиц, объединенных по наличию у них общего признака или вызывающих одинаковые ассоциации. В одну парадигму объединяются языковые единицы, которые могут быть поставлены в соответствие одному объекту или явлению: значению, ситуации, слову, классу слов и т.п.

Часто термин «парадигма» используют как синоним термина «морфологическая парадигма», хотя в зависимости от языкового уровня, к которому относится выделяемый класс единиц, говорят о

- морфологической парадигме,
- синтаксической парадигме,
- лексической парадигме,
- словообразовательной парадигме.

Морфологическая парадигма отражает реализацию грамматических категорий и характеризуется наличием инвариантной части (корня, основы), конечным перечнем грамматических значений и связанных с ними специальных формантов (например, окончаний). Парадигмы слов одной части речи имеют одинаковое внутреннее

устройство и одинаковый набор окончаний. Морфологическая парадигма обычно изображается как таблица форм, устанавливающая соответствие между грамматическими значениями и средствами их выражения. Например, для слова «завод» морфологическая парадигма будет такой:

Грамматические значения (число, падеж)		Флексии
Ед.	И.	_
	Р.	а
	Д.	у
	В.	_
	Т.	ом
	П.	е
Мн.	И.	ы
	Р.	ов
	Д.	ам
	В.	ы
	Т.	ами
	П.	ах

Синтаксическая парадигма — ряд структурно различающихся, но семантически соотносительных синтаксических конструкций — предложений или словосочетаний, связанных в силу их семантической близости отношениями перифразы. Например,

Маша передала Пете книгу.

Пете передана книга от Маши.

Книга передана Машей Пете.

Книга передана от Маши к Пете.

Лексическая парадигма объединяет грамматически однородные слова, имеющие семантическую общность. Например, слова-синонимы, или *«утро — завтрак — будильник — кофе — восход»*.

Словообразовательная парадигма объединяет однокоренные слова, имеющие одну и ту же производящую основу и находящиеся на одной и той же ступени словопроизводства. Например, *делать, переделать, сделать, делающий,…; дело, деловой…*

Синтагматика — раздел науки о языке, занимающийся изучением синтагматических отношений между знаками языка, возникающих между последовательно расположенными его единицами при их непосредственном сочетании друг с другом в реальном потоке речи или в тексте.

Синтагматика изучает отношения между единицами языка «по горизонтали», в отличие от парадигматики, изучающей отношения между единицами языка «по вертикали».

Синтагматические отношения непосредственно наблюдаемы и основаны на линейном характере речи и свойстве ее протяженности, однонаправленности, последовательности. Элементы языка, следуя один за другим, образуют определенные языковые цепочки — **синтагмы**, внутри которых составляющие их элементы вступают в синтагматические отношения.

Поскольку почти все языковые единицы находятся в зависимости либо от того, что их окружает в потоке речи, либо от тех частей, из которых они состоят сами, развитие процедур синтагматического анализа идет по двум направлениям: а) валентностный

анализ и — шире — анализ сочетаемости языковых единиц и б) дистрибутивный анализ.

В широком смысле в языкознании под валентностью понимается общая сочетательная способность слов и единиц иных языковых уровней. В узком смысле понятие валентности сопоставимо с понятием n-местного предиката в логике предикатов.

Дистрибутивный анализ — метод исследования языка, основанный на изучении окружения (дистрибуции, распределения) отдельных единиц в тексте и не использующий сведений о полном лексическом или грамматическом значении этих единиц.

Процедуры синтагматического анализа реализуют прежде всего приемы членения языковых последовательностей и определения их состава, а также особые способы обнаружения влияния одной единицы на другую или их взаимодействия. Особенно четко это проявляется в фонологии и морфологии.

Парадигматические отношения, в отличие от синтагматических, не линейны и не одновременны в потоке речи или тексте, они представляют собой соотношения между элементами языка, объединяемыми в сознании говорящего некими ассоциациями (в силу общности их формы и/или содержания). В случае парадигматических отношений присутствие одного из членов парадигматического ряда в синтагматической цепочке исключает наличие другого, но делает возможной их взаимозамену. Т.е. синтагматические отношения проявляются в совместной встречаемости языковых единиц, а парадигматические — в их взаимоисключении и взаимозамене.

Синтагматика организована по принципу логической конъюнкции, отношения «и–и», парадигматика же — по принципу логической дизъюнкции, отношения «или–или». Первый тип отношений характеризует речь, процесс, второй — систему языка. Одна и та же сущность, входя в систему языка, осуществляет функцию дизъюнкции, но, входя в текст, проявляет функцию конъюнкции. В тексте языковые единицы сосуществуют, в системе образуют парадигмы.

2. Морфологический компонент лингвистического процессора ЕЯ

Морфологический компонент ЛП — комплекс программ, обеспечивающих морфологический анализ и синтез лексем ЕЯ.

Морфология (словоизменение) — раздел науки, изучающий части речи, их категории и формы слов.

Морфема — минимальная значащая часть слова (корень, приставка, суффикс, окончание, постфикс).

Основа — часть слова без окончания (постфикса).

Флексия — окончание (постфикс).

Лексема — слово, рассматриваемое как единица словарного состава языка в совокупности его конкретных грамматических форм и выражающих их флексий, а также возможных конкретных смысловых вариантов.

Словоформа — лексема в некоторой грамматической форме.

Морфологическая парадигма — система форм одного слова (обычно задается таблицей).

Омонимия — звуковое совпадение различных языковых единиц, значения которых не связаны друг с другом.

Лексические омонимы — одинаково звучащие и пишущиеся слова, не имеющие общих элементов смысла и не связанные ассоциативно. Например, *лук* (растение) — *лук* (для стрельбы), *flaw* (трещина) — *flaw* (порыв ветра), *брак* (изъян) — *брак* (женитьба). Различаются полная омонимия, когда совпадают все формы слов, и частичная омонимия, при которой совпадают только отдельные формы слов, называемые **омоформами**. Например, *стих* (глагол в прошедшем времени, единственном числе мужского рода) — *стих* (существительное в единственном числе, именительном падеже), *saw* (пила) — *saw* (2-я форма глагола «to see»).

Омографы — слова, имеющие одинаковое написание, но различное произношение. Например, *мук**а**-м**у**ка*, *lead* [led] (свинец) — *lead* [li:d] (вести), *tear* [teə] (рвать) — *tear* [tiə] (слеза).

Омофоны — слова, которые произносятся одинаково, но различаются в написании. Например, *косный – костный*, *write — right*, *week — weak*.

2.1. Морфологическая модель естественного языка.

«Создание модели есть
доказательство ясности понимания»

Существующие в настоящее время морфологические модели различаются в основном по следующим параметрам.

Во-первых, морфологические модели отличаются по результатам работы основанных на них морфологических анализаторов. На вход морфологический анализатор получает словоформу некоторого ЕЯ, а на выходе может выдавать все значения грамматических

характеристик (род, число, падеж, вид, лицо и т.п.) заданной словоформы, а может просто отвечать на вопрос, принадлежит ли заданная словоформа некоторому ЕЯ или нет (в этом случае морфологические анализаторы называют акцепторами).

Во-вторых, морфологические модели могут ориентироваться на полное покрытие лексики (т.е. все лексемы, которые могут обрабатывать программы морфологического уровня находятся в базе данных) или частичное покрытие лексики (морфологическая модель учитывает возможность появления лексемы, не занесенной в базу данных).

В-третьих, морфологические модели различаются по способу представления и членения словоформ. Существует два основных способа представления лексем.

1) В базе данных хранятся все словоформы всех лексем (возможно, с набором их грамматических характеристик), и каким-то образом определяются словоформы, принадлежащие одной лексеме. Такой способ представления лексем удобен и эффективен для малофлективных языков, в которых различные грамматические категории реализуются, в основном, не с помощью вариации флексий, а некоторым грамматическим способом, например, с помощью предлогов. К малофлективным языкам относится, например, английский язык.
2) В базе данных хранятся основы лексем и списки флексий (возможно, с приписанными им значениями грамматических характеристик), которые присоединяются к основе для получения какой-либо словоформы. Такой способ представления лексем эффективен для флективных языков, в которых различные грамматические категории реализуются путем вариации флексий. Флективным является, например, русский язык. Модели, в которых принят данный способ представления лексем подразделяются еще на две группы: в одной учитываются чисто орфографические основы и флексии, в другой — так называемые псевдоосновы (неизменяемая начальная часть слова) и псевдофлексии (варьируемая при словоизменении конечная часть слова). Выбор того или иного варианта определения основы связан, в основном, с эффективностью реализации и назначением морфологического компонента в целом.

В любой морфологической модели, учитывающей значения грамматических характеристик лексем, с каждой лексемой связаны: синтаксический класс (часть речи), словоизменительный (парадигматический) класс и значения грамматических категорий, или грамматических переменных (ГП), соответствующих синтаксическому классу. Различаются свободные и связанные ГП. Связанные ГП — ГП, присущие лексеме в целом (всем ее словоформам), например, одушевленность и род для существительных. Свободные ГП — совокупность ГП, по которым лексема изменяется, например, число и падеж для существительных.

В один синтаксический класс объединяются лексемы, имеющие

— общий набор ГП,
— общий набор свободных ГП,
— общее множество значений ГП,
— общие синтаксические функции.

В грамматике (русского языка) выделяются следующие синтаксические классы, с которыми связаны следующие ГП (для классов неизменяемых лексем ГП не указаны).

— **Существительные**. ГП — одушевлённость, род, число, падеж. Свободные ГП — число, падеж.
— **Прилагательные**. ГП — одушевлённость, род, число, падеж, степень. Свободные ГП для полных форм — одушевленность, род, число, падеж.

Свободные ГП для кратких форм — род, число. Свободные ГП для сравнительной степени — степень.

- **Глаголы**. ГП личных форм глагола - возвратность, вид, наклонение-время, лицо, род, число; кроме того, переходные глаголы имеют формы страдательного залога. Свободные ГП личных форм глагола — наклонение-время, лицо, род, число, залог. Причастия и деепричастия являются глагольными формами и входят в парадигму глагола. ГП причастий — возвратность, вид, время, залог, одушевленность, род, число, падеж. Парадигма причастий совпадает с парадигмой прилагательных, но у причастий нет форм сравнительной степени. Свободные ГП для полных форм причастий — одушевленность, род, число, падеж. Свободные ГП для кратких форм причастий — род, число. ГП деепричастий — возвратность, вид, время. Свободные ГП деепричастий — время. Иногда удобно связать с глагольной лексемой чисто синтаксическую характеристику — переходность.
- **Наречия**.
- **Личные местоимения**. ГП — одушевленность, род, число, падеж, лицо. Свободная ГП личных местоимений — падеж.
- **Предлоги**.
- **Союзы**.
- **Числительные**.
- **Частицы**.
- **Междометия**.
- **Предикативы**.
- **Вводные слова**.

Иногда в морфологических моделях выделяются синтаксические подклассы лексем, имеющие определенные морфологические и/или синтаксические особенности. Например, в русском языке в классе прилагательных можно выделить местоименные прилагательные (*«который»*), притяжательные прилагательные (*«дядин»*), порядковые числительные (*«второй»*).

2.2. Некоторые особенности и закономерности морфологии русского языка.

В парадигме существительных (кроме существительных с неопределенным родом) и прилагательных единственного числа мужского и среднего рода, а также любых существительных и прилагательных множественного числа форму винительного падежа (В.п.) можно определить т.о.:

- форма В.п. одушевленных существительных мужского рода единственного числа совпадает с формой родительного падежа (Р.п.);
- форма В.п. неодушевленных существительных мужского рода единственного числа совпадает с формой именительного падежа (И.п.);
- форма В.п. всех существительных среднего рода единственного числа совпадает с формой И.п.;
- форма В.п. одушевленных существительных любого рода множественного числа совпадает с формой Р.п.;
- форма В.п. неодушевленных существительных любого рода множественного числа совпадает с формой И.п.

В парадигме всех существительных и прилагательных женского рода единственного числа форма предложного падежа всегда совпадает с формой дательного падежа.

В морфологической модели русского языка необходимо учесть наличие неизменяемых существительных, т.е. существительных, у которых все формы совпадают (например, *«кофе»*, *«метро»*).

Почти в каждом склоняемом или спрягаемом синтаксическом классе существуют лексемы, у которых не существуют некоторые формы соответствующей парадигмы (например, существительное *«ножницы»* не имеет форм единственного числа, прилагательное *«рад»* не имеет полных форм). Такая морфологическая особенность должна быть учтена в морфологической модели.

Прилагательные русского языка имеют две сравнительные степени сильную (*«краснее»*) и слабую (*«покраснее»*), которая образуется путем прибавления и флексии сравнительной степени, и префикса.

Самое большое количество форм имеют переходные глаголы несовершенного вида со следующими значениями ГП (на примере глагола *делать*):

- инфинитив (делать),
- настоящее время, ед. число, 1 лицо (делаю),
- настоящее время, ед. число, 2 лицо (делаешь),
- настоящее время, ед. число, 3 лицо (делает),
- настоящее время, мн. число, 1 лицо (делаем),
- настоящее время, мн. число, 2 лицо (делаете),
- настоящее время, мн. число, 3 лицо (делают),
- прошедшее время, ед. число, мужской род (делал),
- прошедшее время, ед. число, женский род (делала),
- прошедшее время, ед. число, средний род (делало),
- прошедшее время, мн. число (делали),
- повелительное наклонение, ед. число, 2 лицо (делай),
- повелительное наклонение, мн. число, 2 лицо (делайте),
- действительное причастие настоящего времени (делающий),
- страдательное причастие настоящего времени (делаемый),
- действительное причастие прошедшего времени (делавший),
- страдательное причастие прошедшего времени (деланный),
- деепричастие настоящего времени (делая),
- деепричастие прошедшего времени (делав/делавши),
- возвратные формы (с –ся/–сь): инфинитив (делаться); настоящее время, 3 лицо (делается, делаются); прошедшее время (делался, делалась, делалось, делались); действительное причастие настоящего времени (делающийся), действительное причастие прошедшего времени (делавшийся).

У всех непереходных глаголов нет никаких возвратных форм и форм страдательного залога.

У всех глаголов совершенного вида нет никаких форм настоящего времени (но появляются личные формы будущего времени) и страдательных форм.

Существуют еще некоторые подклассы глаголов со своим набором форм (возвратные, многократные, двувидовые и безличные), но в рамках данного задания их можно не рассматривать.

Значения ГП:

ГП	Значение ГП
одушевленность	одушевленность неодушевленность;
род	мужской, женский, средний;
число	единственное, множественное;
падеж	именительный, родительный, второй родительный, дательный, винительный, творительный, предложный, второй предложный;
вид	совершенный, несовершенный;
лицо	первое, второе, третье;
залог	действительный, страдательный;
возвратность	возвратность, невозвратность;
время	настоящее, прошедшее;
наклонение-время	настоящее, будущее, прошедшее, сослагательное, повелительное, инфинитив;
степень	сильная, слабая.

2.3. Морфологическая база данных

Морфологическая база данных должна содержать всю информацию, необходимую для работы процедур морфологического анализа и синтеза.

Если в выбранной морфологической модели принят словарь словоформ, то база данных должна содержать все словоформы учитываемых лексем с указанием их грамматических характеристик и принадлежности определенной лексеме.

Если же в морфологической модели принят словарь основ (псевдооснов), то база данных помимо основ учитываемых лексем должна содержать словарь списков флексий (псевдофлексий), соответствующих каждому парадигматическому классу. С каждой флексией должен быть связан набор значений ГП, приписываемый основе с данной флексией. Если в морфологической модели учитываются какие-либо типичные особенности словоизменения (например, чередование букв в основе), то информация о них также должна храниться в базе данных.

Морфологическая БД помимо лексем с регулярным словоизменением должна содержать лексемы с отсутствующими формами (*«ножницы»*, *«рад»*), с супплетивными формами (*«лучше»* для прилагательного *«хороший»*), неизменяемые существительные (*«метро»*). Кроме того, БД обязательно должна содержать омонимичные лексемы (с полной и частичной омонимией).

2.4. Морфологические анализаторы и синтезаторы ЕЯ

На вход программе морфологического анализа поступает словоформа.

Если программа работает со словарем словоформ, то задача морфологического анализа сводится к задаче поиска заданной словоформы в базе данных, где с каждой словоформой связаны ее грамматические характеристики. Если словоформа в словаре находится, то результатом морфологического анализа будут являться приписанные ей грамматические характеристики и начальная форма исходной лексемы, если же словоформа в словаре не находится, значит она не принадлежит выбранному подмножеству лексического состава ЕЯ.

Если же программа работает со словарями основ и флексий или псевдооснов и псевдофлексий, то имеет смысл все равно поискать исходную словоформу в словаре (она будет найдена, если соответствующая лексема неизменяемая или если данная словоформа имеет пустую флексию).

Если словоформа в словаре не нашлась, то можно, например, отщепить от нее последнюю букву (предположительно таким образом поделив словоформу на основу и флексию) и поискать оставшуюся часть в словаре. Если поиск опять оказался неуспешным, нужно отщепить две последние буквы и т.д. Процесс завершается, когда произведен поиск последнего варианта расщепления словоформы на основу и флексию. А это происходит, либо когда отщеплено столько последних букв, какова максимальная длина флексии (с учетом постфикса, например, –ся/–сь) в соответствующем ЕЯ (в случае словаря основ и флексий), либо когда проанализирована пустая основа и вся словоформа как флексия (в случае словаря псевдооснов и псевдофлексий).

При этом, если ни один из вариантов основы или псевдоосновы в словаре не находится, то заданная словоформа не принадлежит выбранному подмножеству лексического состава ЕЯ. Если же какой-либо (или какие-либо, например, для словоформы *«дома»* — *«дом-**а**»* (сущ.) и *«дома»* (наречие)) вариант предположительной основы нашелся в словаре, надо проверить, может ли у данной основы быть предположительная флексия. Если да — результатом анализа словоформы являются грамматические характеристики, связанные с флексией и начальная форма соответствующей лексемы, если нет — надо продолжить процесс расщепления исходной словоформы на основу и флексию.

Отметим, что при использовании любого словаря результат морфологического анализа в общем случае неоднозначен в силу наличия в ЕЯ морфологической омонимии.

На вход программе морфологического синтеза поступают: а) лексема в начальной форме и б) значения свободных грамматических переменных (в некотором заранее оговоренном виде), соответствующих запрашиваемой словоформе данной лексемы или запрос на синтез всех форм заданной лексемы.

Результатом работы программы морфологического синтеза является либо словоформа с запрашиваемыми грамматическими характеристиками, либо все формы заданной лексемы. Морфологический синтез также может оказаться неоднозначным в случае вариативности флексии в какой-либо форме слова или при морфологической омонимии.

3. Синтаксический компонент лингвистического процессора ЕЯ

Синтаксис — раздел грамматики, изучающий процессы порождения речи: сочетаемость и порядок следования слов внутри предложения, а также общие свойства предложения как автономной единицы языка и высказывания как части речи.

Одним из центральных дискуссионных вопросов в области компьютерной лингвистики является вопрос о том, каковы задачи и место синтаксического этапа анализа в процессе определения смысла текста: речь идёт прежде всего о соотношении синтаксического и семантического уровней анализа и вообще о целесообразности разделения этих уровней в модели понимания ЕЯ. Существуют два принципиально различных подхода: модульный и интегральный.

Системы модульного типа. В этих системах каждому уровню лингвистического анализа соответствует отдельный компонент системы. Системы модульного типа допускают разные схемы взаимодействия компонентов (последовательная работа, параллельный перемежающийся анализ). Это не меняет существа дела: синтаксис и семантика обрабатываются в системе разными механизмами. При этом синтаксический уровень понимания входного текста выделен в отдельный блок, преобразующий текст в его синтаксическое представление.

Системы интегрального типа. В таких системах синтаксический и семантический анализаторы (а часто и анализатор прагматического уровня) слиты в отдельный блок. Система ориентируется сразу на формирование (на основе текста) достаточно богатых концептуальных структур, а не на постепенную «глубинизацию» понимания, как это имеет место в системах модульного типа. Здесь не предусматривается формирование синтаксического представления входного текста. Синтаксическая информация используется фрагментарно и лишь как вспомогательная.

Системы интегрального типа успешно применяются пока только в ЕЯ-системах, работающих в предельно узкой проблемной области, и остается открытым вопрос о том, насколько они эффективны при обработке больших и разнообразных по тематике массивов текстов. Наиболее интересными и перспективными представляются системы модульного типа. И в данной работе рассматриваются, в основном, именно такие системы.

Среди сторонников систем модульного типа также нет полного единодушия, например, в вопросе о том, насколько развитым и «семантизированным» должен быть синтаксический этап анализа. Это находит отражение в разной степени дифференцированности синтаксических отношений, в разной глубине интерпретации синтаксических отношений, а также в широте привлечения семантической информации при построении синтаксической структуры входного предложения.

3.1. Синтаксическая модель естественного языка

При создании синтаксического компонента необходимо разработать **синтаксическую модель** соответствующего ЕЯ, для чего необходимо определить следующее: способ описания синтаксиса языка, способ представления синтаксической структуры предложения, метод анализа и метод синтеза предложений на ЕЯ.

3.2. Модели представления синтаксической структуры предложения

3.2.1. Деревья зависимостей

Деревья зависимостей — наиболее наглядный и наиболее распространенный способ представления синтаксической структуры предложения. При этом предложение представляется как линейно упорядоченное множество элементов (словоформ), на котором можно задать ориентированное дерево (узлы — элементы множества). Каждая дуга, связывающая пару узлов, интерпретируется как *подчинительная* связь между двумя элементами, направление которой соответствует направлению данной дуги.

Множество всех узлов дерева, прямо или косвенно зависящих от какого-либо узла, включая сам этот узел, составляет группу зависимости этого узла.

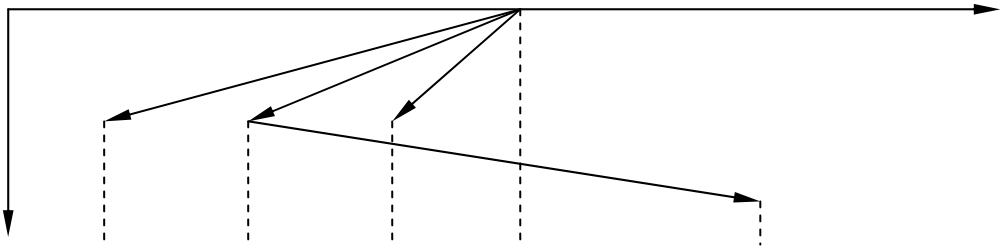
Проективность — важное свойство древовидных структур, отражающее связь между отношением линейного порядка и отношением подчинения. Деревья зависимостей называются проективными, если для любого узла группа зависимости этого узла является неразрывным отрезком предложения.

Проективность предложения легко определяется при графическом изображении дерева зависимостей. При этом на плоскости рисунка выбирается прямоугольная система координат (ось ординат направлена сверху вниз). Узлы дерева (слова предложения) изображаются целочисленными точками плоскости: абсцисса узла — порядковый номер слова в предложении, ордината — высота слова в дереве. При таком способе изображения предложение проективно, если дуги дерева не пересекаются с вертикалями, проведенными из узлов (сверху вниз), и между собой. Различаются проективные и слабопроективные деревья зависимостей.

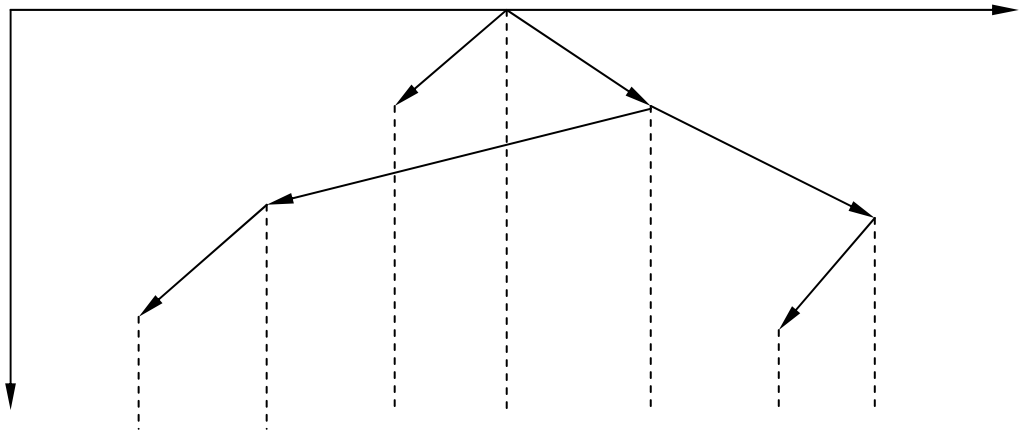
Дерево зависимости проективно, если для любых трех его узлов ***a***, ***b***, ***c*** из того, что ***b*** зависит от ***a***, и ***c*** лежит между ними, следует, что ***c*** зависит от ***a*** или от ***b***.

Дерево зависимости слобопроективно, если для любых его четырех узлов ***a***, ***b***, ***c***, ***d*** из того, что ***b*** зависит от ***a***, и ***d*** зависит от ***c*** следует, что пары ***a***, ***b*** и ***c***, ***d*** не разделяют друг друга (т.е. любые два интервала — при линейном изображении дерева зависимостей — либо не пересекаются, либо один из них содержится в другом).

Пример непроективного предложения: *«Я памятник себе воздвиг нерукотворный»*.



Пример слабопроективного предложения: *«Этому человеку мы будем обязаны всю жизнь»*.



Этому человеку мы будем обязаны всю жизнь

В деловой прозе деревья зависимостей подавляющего большинства предложений проективны. За исключением некоторых особых случаев непроективность предложений в деловом тексте — верный признак недостаточной грамотности автора (хотя в художественной литературе, особенно в поэзии, отклонения от проективности вполне обычны).

Недостатки способа представления синтаксических структур в виде деревьев зависимостей:

1) жесткое требование рассматривать каждое формально выделенное вхождение слова в качестве отдельного элемента предложения;
2) все без исключения связи между словоформами трактуются как подчинительные.

3.2.2. Структуры непосредственно составляющих (НС-структуры)

НС-структура — множество отрезков предложения, называемых *составляющими*, которое удовлетворяют следующим условиям:

— в качестве элементов множества отрезков предложения присутствуют само предложение и все его отдельные словоформы;
— в одну составляющую объединяются отрезки непосредственно синтаксически связанные между собой;
— любые две составляющие либо не пересекаются, либо одна из них содержится в другой.

С помощью НС-структур в предложении можно выделить не только отдельные слова, но и некоторые словокомплексы, функционирующие как единое целое (например, *«будем обязаны»*). С помощью НС-структур более естественно описываются конструкции с неподчинительными отношениями.

Пример НС-структуры (для наглядности каждая словоформа не заключается в скобки):

((Английские колонизаторы) доводили их (до (нищеты, (непрекращающегося голода) и (чудовищного (роста смертности)))))

Недостатки НС-структур:

1) неоднозначность трактовки силы связи между элементами словосочетаний приводит к неоднозначным НС-структурам; например, *((чудовищного роста) смертности)* или *(чудовищного (роста смертности))*.
2) НС-структуры не вводят никакой иерархии среди составляющих одного уровня.
3) Невозможно представить непроективные предложения.

3.2.3. Ориентированные структуры непосредственно составляющих (ОНС-структуры)

ОНС-структура — это структура составляющих, где для каждой неодноэлементной составляющей определена одна из её НС в качестве главной (неглавные зависят от главной).

Всякая ОНС-структура однозначно определяет соответствующее ей дерево зависимостей или НС-структуру (обратное неверно).

ОНС-структуры наследуют недостаток деревьев зависимостей — неспособность адекватно описывать неподчинительные связи.

3.2.4. Частично ориентированные структуры непосредственно составляющих (ЧОНС-структуры)

Различия в формальных определениях ОНС-структур и ЧОНС-структур заключается в том, что главные НС выделяются не для всех элементов множества неодноэлементных составляющих, а лишь для некоторого его подмножества.

ЧОНС-структуры дают возможность описывать как подчинительные, так и неподчинительные связи, что существенно не только для представления конструкций с однородными членами, но и для более точного отражения отношений между компонентами аналитических сочетаний, например, форм составного будущего времени (*«буду читать»*) или предложно-падежных сочетаний (*«в школе»*).

3.3. Типы формальных грамматик, используемых для описания синтаксиса естественного языка

Описание основных типов формальных грамматик можно найти в [10], [11].

Перечисленные выше способы представления синтаксической структуры предложения на ЕЯ описываются соответственно следующими формальными грамматиками.

3.3.1. Грамматика зависимостей (G_D)

$G_D = \langle V_T, V_N, V_S, R_T, R_N \rangle$, где

- V_T — алфавит терминальных символов,
- V_N — алфавит нетерминальных символов — классов терминалов,
- V_S — множество корневых классов, $V_S \in V_N$.
- R_T — множество правил классификации вида $A \rightarrow a$ (терминал a принадлежит классу A),
- R_N — множество правил кустов вида $A(B_1 \quad B_k * B_{k+1} \quad B_n)$ или $A(*)$, которые для каждого класса A задают его систему управлений (классами B_j), выраженную в терминах классов, с указанием их линейного порядка относительно корня куста и друг друга.

Язык, порождаемый грамматикой зависимостей, — это множество терминальных цепочек $a_1 \ldots a_n$, где каждый символ a_i можно отнести к определенному классу A_i, и для любого A_i в грамматике существует соответствующее правило куста $r \in R_N$.

3.3.2. Контекстно-свободные грамматики (G_{CF})

Описание G_{CF} см. в [10], [11]**.**

Вывод каждой цепочки в G_{CF} можно изобразить в виде дерева. Множество поддеревьев дерева соответствует множеству непосредственно составляющих порождаемой цепочки. Метка корня дерева — название полной составляющей предложения, а метки узлов-сыновей — имена соответствующих непосредственно составляющих.

3.3.3. Ориентированные контекстно-свободные грамматики $\langle G_{CF}, \Delta \rangle$

Δ — ориентировка грамматики G_{CF}, которая вводится следующим образом: из множества правил R выделяется подмножество R^1, в которое входят все правила вида $A \rightarrow \alpha_1 \ldots \alpha_n$ при $n \geq 1$; $\alpha_1, \ldots \alpha_n \in V_G$. Для каждого из этих правил в цепочке $\alpha_1 \ldots \alpha_n$ маркируется одно из вхождений α_k в качестве главного (например, сверху *). Выделенный элемент может быть как терминальным, так и нетерминальным.

3.3.4. Частично ориентированные контекстно-свободные грамматики $\langle G_{CF}, \Delta' \rangle$.

Отличие частично ориентированных контекстно-свободных грамматик от ориентированных контекстно-свободных грамматик заключается в том, что частичная ориентировка Δ' вводится не на всем множестве R^1, а на некотором его подмножестве.

3.3.5. Сетевые грамматики

Сетевые грамматики представляют собой одновременно аппарат для описания системы языка и для задания процедуры анализа предложений на основе понятия конечного автомата (см. [10], [11]). Задаются такие грамматики в виде графа (сети переходов).

Грамматика, заданная в виде конечного автомата, неспособна приписывать анализируемым цепочкам внутреннюю структуру. Но внутреннюю структуру можно фиксировать с помощью системы конечных автоматов (СКА), которую можно задать в виде рекурсивной сети переходов. СКА представляет собой совокупность конечных автоматов, среди которых выделен один главный автомат, с которого начинается работа СКА.

Отличие рекурсивной сети переходов от сети переходов заключается в том, что в рекурсивной сети переходов дуги переходов могут быть помечены как терминальными, так и нетерминальными символами (представляющими собой отдельные конечные автоматы).

Результат анализа входной цепочки посредством СКА определяется трассой движения по рекурсивной сети переходов. Трасса — это последовательность всех терминальных и нетерминальных символов, помечающих дуги, по которым совершается проход в процессе анализа заданной цепочки от начального состояния главного автомата до его конечного состояния, при этом символы, раскрывающие какой-либо нетерминальный символ, заключаются в круглые скобки. Естественно, допускается вложенность скобочных структур. СКА эквивалентна контекстно-свободным грамматикам, а содержимое выходной последовательности в момент завершения анализа цепочки представляет собой структуру данной цепочки в терминах непосредственно составляющих. Если в каждом автомате выделить главное состояние, то в результате можно получить ОНС-структуру. Если же главное состояние выделять только в некоторых конечных автоматах, то можно получить ЧОНС-структуру.

СКА присуще общее для всех контекстно-свободных грамматик ограничение — невозможность учета синтагматических свойств языковых единиц, проявляющихся в конкретных контекстуальных условиях. Это может привести к появлению лишних, неправильных структур. Но рекурсивные сети переходов не исчерпывают всех возможностей сетевых грамматик. Наиболее мощной среди сетевых грамматик является модель В. Вудса, названная **расширенной сетью переходов** (РАСП).

РАСП строится на базе рекурсивной сети переходов, но располагает средствами контроля над ходом анализа, состоящими в проверке определенных условий при переходе из одного состояния в другое и выработке некоторых указаний относительно дальнейшего продвижения по сети. Эти средства представляются в виде операторов, указанных на дугах. Операторы выполняют роль фильтров.

Один из примеров практического использования РАСП для синтаксического анализа английского языка описан в [12]. В этой работе РАСП представлена в несколько нетрадиционной форме и дополнена рядом элементов, позволяющих удобно и эффективно реализовать алгоритм синтаксического анализа.

Описание формальной грамматики ЕЯ представляет собой набор иерархически организованных грамматических сетей переходов. Каждая сеть строится по следующим правилам:

```
<грамматическая сеть>::= <имя сети> <список глобальных полей сети>
                          (<список состояний перехода>)
                          <список локальных полей сети> <список процедур>
          <имя сети> ::= $<строка>
<список глобальных полей сети> ::= {<строка(имя)>}
 <список локальных полей сети> ::= {<строка(имя)>}
             <список процедур> ::= {@<строка(имя)>}
   <список состояний перехода> ::= <состояние сети>{;<состояние сети>}
    <состояние сети> ::= <номер состояния> : <список альтернатив>
   <номер состояния> ::= <целое без знака>
<список альтернатив> ::= <альтернатива> {<альтернатива>}
      <альтернатива> ::= <элемент перехода> <точка перехода> <список процедур>
  <элемент перехода> ::= <имя сети> |"<лексема>" | <обобщенная лексема>
    <точка перехода> ::= <номер состояния> | *
           <лексема> ::= <строка>
<обобщенная лексема> ::= <грамматическая информация соотв. класса лексем>
```

Обход сети начинается с состояния с номером «0», которое обязательно должно присутствовать в каждой сети. При этом последовательно просматриваются все альтернативы состояния. Вообще говоря, сколько различных альтернатив в состоянии, столько различных вариантов возможных продолжений построения синтаксической конструкции, описанной данной сетью. Алгоритм обхода сети позволяет реализовать многовариантный синтаксический анализ исходной фразы (и любой синтаксической конструкции, в частности).

Если текущая подцепочка входной цепочки словоформ «соответствует» очередной альтернативе (т.е. либо удалось свернуть некоторую вложенную сеть, либо первая словоформа исследуемой подцепочки совпала с указанным элементом альтернативы), то происходит переход в состояние, номер которого указан в альтернативе (звездочкой (*) обозначается точка выхода из сети). Затем процесс повторяется, исходя из нового состояния, до тех пор, пока либо ни один элемент альтернативы активного состояния не окажется подходящим (т.е. анализируемый фрагмент фразы не удовлетворяет данному пути описания ожидаемой синтаксической конструкции), либо очередным состоянием перехода будет * (что означает успешный вариант свертки по текущей сети).

Проходя по той или иной сети, можно свернуть (выделить) ту или иную синтаксическую конструкцию исходной фразы: простое предложение, именную группу, предложную группу, детерминант существительного, фразовый глагол и прочие.

В процессе свертки синтаксической конструкции могут определиться такие её грамматические характеристики, которые потребуются при включении данной конструкции в более сложные, объемлющие (например, число именной группы, форма глагольной конструкции и другие). Значения этих характеристик запоминаются в глобальных полях текущей сети посредством процедур, возможно перечисленных при альтернативах, и доступны объемлющим сетям.

Если же при свёртке текущей синтаксической конструкции анализируются грамматические характеристики вложенных синтаксических групп (сетей) или конкретных лексем, то их значения можно запомнить в локальных полях сети. Действия, которые необходимо выполнить над локальными полями при выходе из сети (окончании свертки), описаны в процедурах, имена которых перечислены после имен локальных полей в описании сети. Результат выполнения этих процедур также может быть зафиксирован в глобальных полях сети. В качестве примера можно привести процедуру, проверяющую соответствие детерминанта определяемому существительному (по числу существительного) или процедуру, приписывающую множественное число однородной именной группы, состоящей из именных групп единственного числа.

Другими словами все локальные свойства любой синтаксической конструкции анализируются при ее свёртке и забываются, глобальные же характеристики сохраняются для объемлющих сетей.

Такая иерархическая организация грамматики позволяет сворачивать (анализировать) как целую фразу, так и любой ее фрагмент. Можно проверить, например, является ли исходная фраза правильным предложением английского языка, а можно выделить все потенциально возможные правильные именные группы исходной фразы (конечно же, изменив порядок обращения к сетям и предъявление исходных словоформ).

3.4. Синтаксическая база данных

Синтаксическая база данных должна содержать:

- формальное описание грамматики некоторого фиксированного подмножества выбранного ЕЯ;
- описание синтаксических характеристик отдельных лексем или словосочетаний выбранного подмножества ЕЯ (синтаксический класс, синтаксический подкласс, переходность…); все учитываемые синтаксические характеристики могут содержаться в используемой для целей синтаксического анализа морфологической базе данных, в этом случае необходимо иметь программные средства, позволяющие извлекать их оттуда;
- описание моделей управления лексем выбранного подмножества ЕЯ (при соответствующем выборе метода синтаксического анализа).

3.5. Синтаксические анализаторы фраз ЕЯ

Построить синтаксический анализатор ЕЯ значительно сложнее, чем морфологический по ряду причин: нет достаточно четкой и формальной лингвистической литературы, описывающий какой-либо ЕЯ, грамматика естественного языка принципиально

недетерминирована и неоднозначна, синтаксис ЕЯ весьма разнообразен, сложен и произволен (особенно в разговорной речи и в поэзии). Трудными для автоматической обработки являются такие вполне допустимые в ЕЯ явления, как *эллипсис* (пропуск обязательных фрагментов предложения в силу возможности их восстановления из предыдущего контекста: *«Маше нравился Саша. Она — ему.»*), *парцелляция* (разбиение одного грамматического предложения на несколько предложений для усиления акцента на некоторые его фрагменты: *«Приказано нам готовиться. К походу.»*). Некоторые сложные явления языка часто обрабатываются специальными процедурами до работы синтаксического анализатора (т.е. осуществляется некоторый предсинтаксический анализ). К таким процедурам можно отнести, например, процедуры обрабатывающие фразиологизмы, группу числительного, проверяющие правильность расстановки скобок, знаков пунктуации и, возможно, проводящие некоторую дополнительную фрагментацию предложения. Кроме того, само автоматическое разбиения текста на ЕЯ на отдельные предложения является не совсем тривиальной задачей и выполняется на этапе предсинтаксического анализа.

Синтаксические анализаторы различаются между собой следующим: типом анализируемых текстов (деловая проза, художественная литература…); наличием и характером ограничений, накладываемых на структуру анализируемых предложений; наличием требования правильности анализируемой цепочки словоформ; возможностью анализировать только отдельное предложение (или часть предложения) или некоторый фрагмент текста, состоящий более чем из одного предложения; стратегией анализа.

В настоящее время можно говорить о трех основных стратегиях, логико-алгоритмических подходах к построению синтаксических анализаторов.

3.5.1. Стратегия недетерминированного, фильтрового анализа

Процедура синтаксического анализа на первом этапе порождает заведомо избыточный набор синтаксических связей (например, с помощью какой-либо порождающей грамматики), из числа которых на втором этапе с помощью серии фильтров (например, проверка правил согласования) отбираются только те синтаксические структуры входного предложения, которые являются правильными с точки зрения выбранных фильтров. В настоящее время такая стратегия имеет разновидности, которые различаются

а) степенью ослабления контекстных условий на этапе порождения связей;

б) характером применяемых фильтров;

в) статусом синтаксических структур, подвергающихся фильтрации (синтаксическая структура входного предложения, синтаксические структуры фрагментов входного предложения).

Как правило, основанные на такой стратегии анализаторы затрачивают много времени на порождение и фиксацию в памяти ЭВМ избыточных синтаксических структур, которые затем, на этапе фильтрации, будут отвергнуты. Вместе с тем эта стратегия в большей степени, чем другие, гарантирует полноту анализа многозначного предложения.

3.5.2. Стратегия, опирающаяся на механизм возвратов (backtracking)

Отличие данной стратегии от предыдущей заключается в том, что алгоритм на каждом шаге выбирает одну из возможных интерпретаций, но при этом сохраняется принципиальная возможность порождения альтернативных интерпретаций в случае какой-либо неудачи с первой (например, если полученная синтаксическая структура

непроективна, не проходит семантический фильтр и т.п.). При этом анализ предложения прекращается после нахождения первого приемлемого варианта разбора.

Если приемлемый вариант разбора не удается получить одним из первых, то данная стратегия становится похожей на предыдущую. В среднем скорость работы анализатора, опирающегося на механизм возвратов, выше.

Чтобы избежать общего недостатка описанных двух стратегий (перебор большого количества лишних вариантов установления синтаксических связей), в некоторых синтаксических анализаторах применяются различные эвристические методы, управляющие процессом анализа, которые могут позволить получить предпочтительный вариант разбора первым. В качестве эвристик могут быть использованы, например, следующие предпочтения: значения омонимичных лексем можно упорядочить по вероятности их появления в тексте, и в первую очередь можно рассматривать наиболее вероятный вариант, затем (если первый почему-то не подошёл) следующий и т.д.; можно указать наиболее предпочтительные позиции расположения дополнений по отношению к сказуемому. Например, для известного примера Л.В. Щербы *«Глокая куздра штеко будланула бокра и кудрячит бокренка»*, если выбирать наиболее вероятный вариант синтаксической интерпретации первых четырех слов, то получим следующее: *кто — куздра*, куздра *какая — глокая*, куздра *что сделала — будланула*, будланула *как — штеко*. Но возможны также и другие варианты: *кто — куздра*, куздра *какая — глокая*, куздра *что сделала — будланула*, куздра *чья — штеко* или *кто — штеко*, штеко *что сделала — будланула*, будланула *как — глокая* (деепричастие), глокая *кого — куздра*. Правда, в последнем варианте должен быть отмечен пропуск запятой после деепричастного оборота.

3.5.3. Стратегия детерминированного анализа

Алгоритм синтаксического анализа работает таким образом, что ни одна синтаксическая связь, установленная в процессе анализа предложения не может в последствии быть отвергнута, т.е. она обязательно присутствует в одной из синтаксических структур, являющихся результатом работы синтаксического анализатора.

При использовании стратегии детерминированного анализа вся языковая информация, которая в принципе может повлиять на установление связи между синтаксическими единицами предложения, привлекается одновременно. Причем, при установлении каждой связи должны соблюдаться такие условия, которые гарантировали бы получение связной синтаксической структуры предложения на выходе. Т.е. для окончательного вывода о наличии связи между двумя синтаксическими единицами необходимо проверить (кроме условий на сочетаемость) некоторые контекстные условия (наличие или отсутствие в фиксированной позиции других синтаксических единиц с заданными характеристиками, наличие или отсутствие в фиксированной позиции тех или иных знаков препинания и т.п.). Набор таких условий, сформулированных, для больших классов пар синтаксических единиц, описывает *синтаксическую ситуацию*, диагностичную для расстановки связей.

В основе стратегии детерминированного анализа лежит инвентарь синтаксических ситуаций, которые учитываются выбранной моделью синтаксиса ЕЯ. Описание каждой ситуации может быть задано декларативно или в процедурном виде — это зависит от языка программирования. Каждая синтаксическая ситуация привязана к какому-либо грамматическому явлению: наличие в предложении однородных членов, наличие причастного или деепричастного оборота, наличие конкретной грамматической формы подлежащего или сказуемого и т.п.

В целом, стратегия детерминированного анализа ориентирована на однозначный грамматический разбор (и в этом его слабое место). Однако, не исключены ситуации, в

которых синтаксический анализатор не имеет достаточной информации для однозначного выбора. Тогда либо все-таки как-то выбирается один из вариантов грамматического разбора, либо строятся несколько альтернативных вариантов.

Анализаторы, основанные на стратегии детерминированного анализа, являются достаточно быстродействующими и эффективными. Однако, для достижения эффективности синтаксического анализа произвольных (даже только синтаксически правильных) предложений ЕЯ требуется создать адекватный и полный инвентарь синтаксический ситуаций, что крайне трудоемко и принципиально сложно.

3.6. Синтаксические отношения (связи, зависимости)

Наборы синтаксических отношений в разных синтаксических моделях различны и отличаются степенью дифференцированности и уровнем интерпретации. Например, существуют модели, в которых различаются только сочинительные и подчинительные связи без их дальнейшей дифференциации. Особенности используемого набора синтаксических отношений зависят, в частности, от того, предусмотрен ли дальнейший семантический анализ, каковы его функции и способ взаимодействия с синтаксическим анализатором.

Среди множества синтаксических отношений выделяется особая группа, соответствующая *актантным отношениям предикатного слова*.

Предикат — слово, подчиняющее себе другие слова и синтаксические конструкции предложения и определяющее их грамматическую форму, а иногда и значение.

Предикат можно рассматривать как фрейм (шаблон): сам предикат — имя фрейма (вершина фрейма), а подчиненные ему синтаксические конструкции — слоты фрейма (валентности предиката). Для каждой валентности предиката определены условия заполнения этой валентности (значения грамматических характеристик, семантическое значение) и конкретное синтаксическое отношение. Синтаксические отношения часто задаются с помощью вопросительного слова (*кто?*, *куда?*, и т.п.). Совокупность синтаксических отношений, задаваемых предикатом (фрейм предиката), часто называют **моделью управления предиката**. Отметим, что у одного предиката может быть несколько разных моделей управления.

Актант — слово или синтаксическая конструкция, заполняющая валентность предиката.

Предикатами в русском языке являются глаголы, глагольные формы, отглагольные существительные и прилагательные и предлоги.

Например, описание модели управления для предлога *к* (*предложной группы предлога к*) может выглядеть так:

к → [куда?, к кому/чему?] существительное с зависящими от него словами и конструкциями (группа существительного): одушевленность – любая, род – любой, число – любое, падеж - дательный;

а для глагола *идти* так:

идти → [кто?] — группа существительного: одушевленное, род – любой, число – любое, падеж – именительный;

[куда?] — предложная группа предлога в (существительное неодушевленное), предлога к или предлога на (существительное неодушевленное);

[откуда?] — предложная группа предлога из (существительное неодушевленное) или предлога от;

Актанты предиката могут быть *обязательными* (т.е. они должны в том или ином виде обязательно присутствовать в предложении, содержащем данный предикат) и *необязательными* (т.е. они могут в реальном предложении отсутствовать). Информацию об обязательности актанта также целесообразно хранить в описании модели управления предиката. Кроме того, полезно указывать и информацию о предпочтительном или обязательном взаимном линейном расположении в тексте предиката и его актантов, о невозможной или желательной сочетаемости актантов.

Модели управления предикатов являются формализованной записью ограничений на грамматические и/или семантические характеристики и, возможно, на способы совместного использования в тексте зависящих от них слов и конструкций. Вообще говоря, в языке у всех слов (не только у предикатов) могут быть зависящие от них другие слова, на которые могут быть наложены соответствующие ограничения (правила сочетаемости отдельных слов и синтаксических групп). Составление таких **обобщенных моделей управления** можно использовать как средство описания ЕЯ.

При использовании МУ в качестве основы описания языка можно достичь произвольной гибкости и детальности, становятся непринципиальными ограничения на степень грамматичности языка, не разделяется явно семантическая и синтаксическая информация. Ожидается, что при наличии МУ, описывающих язык, задачу синтаксического анализа можно считать решенной без каких-либо уточнений, ограничений на входной язык, сферу применимости и прочее. Т.е., задача синтаксического анализа сводится к задаче построения множества моделей управления. Задача эта, безусловно, очень непроста и трудоемка. Кроме того, для эффективности использования обобщенных моделей управления необходимо учитывать их частотные характеристики и контекст (категория текстов, для которой является специфичным употребление определенных слов и грамматических конструкций).

3.7. О многовариантности синтаксического анализа

Принципиальная многовариантность синтаксического анализа — узловая проблема для разработчиков синтаксических анализаторов.

Многовариантность возникает не только в связи с наличием морфологической омонимии (см. выше), но и синтаксической омонимии.

Синтаксическая омонимия — возможность выделения разных смыслов у одного предложения, обусловленная наличием у него разных синтаксических структур.

Примеры предложений, для которых принципиально невозможно разрешить синтаксическую омонимию:

«Мать любит дочь».

«Молодые мужчины и женщины…».

«Письмо отцу друга…».

«Тощая торговка вяленой воблой торчала среди ящиков».

«Сплочение рабочих бригад вызвало осуждение товарища министра».

«Привет освободителям Харькова от немецко-фашистских захватчиков».

«Школьники из Старицы поехали в Торжок».

«Это потрясло до глубины души оскорбленного брата».

«Девочка вытерла тщательно вымытую посуду».

«Я вижу только два дерева».

«Таблица допустимых размеров …».

«Я видел его молодым».

4. Семантический компонент ЕЯ-систем

Семантика — раздел языкознания, изучающий все содержание, информацию, передаваемые языком или какой-либо его единицей.

Понятие — мысль, отражающая в обобщенной форме предметы и явления действительности посредством фиксации их свойств и отношений.

Концепт — понятие.

Значение языкового выражения(ЯВ):

- **синтаксическое** — система, ассоциированных с данным ЯВ эталонных парадигматических, синтагматических и иерархических связей с другими знаками языка;
- **сигматическое** — класс реальных объектов, в соответствие которым может быть поставлено ЯВ;
- **семантическое** — класс эталонных психических моделей реальных объектов (или класс концептов), в соответствие которым может быть поставлено ЯВ;
- **прагматическое** — класс нормативно соотнесенных с ЯВ действий потенциальных реципиентов или же класс действий и целей потенциального автора сообщения, побуждающих его к речевой деятельности.

Смысл ЯВ — соотнесенная с ЯВ в реальном процессе речевой деятельности подсистема значения.

Понимание ЯВ — процесс раскрытия смысла ЯВ реципиентом, т.е. установление тех сторон значения, которые наиболее существенны в текущей ситуации с его точки зрения и которые, как он предполагает, имел в виду автор сообщения.

Однако не всегда смысл, соотнесенный с сообщением реципиентом, совпадает со смыслом, вкладываемым в сообщение автором, а любой их них может не совпадать с наиболее вероятной в языке интерпретацией сообщения (нормативно выделенной подсистемой значения ЯВ), т.е. смыслом относительно языка, критерии выделения которого должны быть объективными, например, синтаксическими. В качестве примеров возможно различного понимания ЯВ автором, реципиентом и относительно языка (объективно) можно привести следующие ЯВ:

Этого просто не вынести!

«А что вам нужно вынести?» — спросила Алиса (Л. Кэрролл);

За безбилетный проезд и провоз одного места багажа взимается штраф… (объявление в общественном транспорте).

4.1. Теория концептуальной зависимости Р. Шенка

Классическим примером экспериментальной системы интегрального типа, в которой подробно исследован семантический аспект ЕЯ и предлагается интересный подход к решению проблемы понимания текста на ЕЯ (английском), является система MARGIE Р. Шенка [8]. В основе MARGIE лежит представление смысла фраз ЕЯ в терминах теории концептуальной зависимости (**ТКЗ**), т.е. оно состоит из понятий, объединенных определенными отношениями между ними. Эта система умеет производить

умозаключения, вытекающие из смысла, заключенного во входном сообщении, и осуществлять перифразирование входных предложений ЕЯ.

4.1.1. Основные положения ТКЗ Р. Шенка

Каждое слово, входящее в текст, рассматривается как *понятие (концепт)*, представляющее собой набор свойств, связанных с ним, часть из которых может быть известна системе, а часть — нет.

Не делается явных различий между лингвистическими и нелингвистическими знаниями.

Чтобы понимать, надо делать *предположения* (возможно, ошибочные), исходя из знаний, хранящихся в системе.

Базовым механизмом восприятия, используемым программой являются *ожидания* — описание ситуации, которая рассматривается как наиболее вероятная в ближайшем будущем.

Концептуальная память системы содержит только понятия (а не слова).

Существуют четыре *концептуальных падежа* (отношений) — *объектный* (O), *директивный* (D), *реципиентный* (R), и *инструментальный* (I).

Сложные понятия, как и смысл всей входной фразы, строятся из менее сложных на основе правил *концептуального синтаксиса*, т.е. правил конструирования отношений между понятиями на концептуальном уровне.

Концептуальные правила используют *концептуальные категории* (типы понятий).

Концептуализация — идея.

Концептуализация может состоять из деятеля, действия и определенного набора концептуальных падежей, а также – из объекта и описания состояния, в котором он находится, или изменения его состояния.

Концептуальные структуры имеют в своей основе *элементарные действия* — АКТЫ (их всего 11!).

АКТы воздействуют на память системы, кроме того, в соответствии с ними могут производиться *умозаключения*. Умозаключение — концептуализация, которая может быть выведена из другой концептуализации с вероятностью меньшей 1.

4.1.2. Концептуальные категории ТКЗ

PP — только физические объекты (одушевленные и неодушевленные). Они могут быть субъектами действия, объектами, играть роль направления и реципиента.

АКТ — действия.

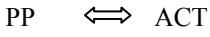
LOC — местоположения. Для каждого физического АКТа оно определяет, где происходит включающая его концептуализация. LOC может модифицировать концептуализацию и выполнять роль направления.

T — времена. Время считается точкой на оси времени. Эта точка может быть абсолютной (6 часов вечера 28 ноября 2005 года) или относительной (вчера).

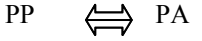
AA — Action aider — модификации свойств акта. Например, параметр скорости для действия PROPEL (см. ниже) есть AA. Таких категорий немного.

PA — атрибут объекта (со своими характеристиками или значениями, например, «цвет», «размер»). Считается, что PP состоят из набора PA, которые их определяют.

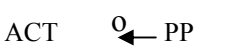
4.1.3. Правила концептуального синтаксиса



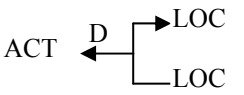
— PP могут производить действия



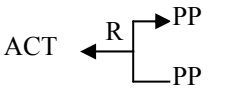
— PP могут описываться через какое-либо свойство



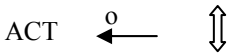
— АКТы имеют объекты



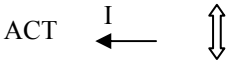
— АКТы имеют направление



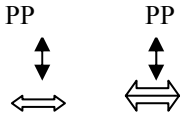
— АКТы имеют реципиентов



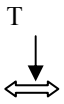
— MTRANS требует в качестве объекта концептуализацию, а MBUILD имеет свой собственный тип объекта (см. ниже)



— АКТы могут иметь концептуализации в качестве инструмента

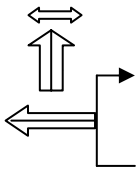


— PP могут быть описаны через концептуализацию, в которой они встречаются инструмента

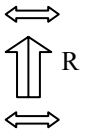


— в концептуализациях присутствует время

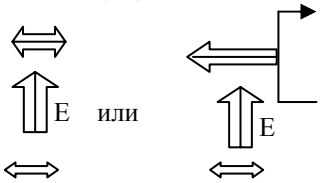
— концептуализации имеют местоположения



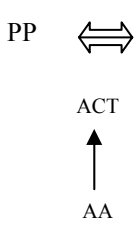
— концептуализации могут иметь результатом изменение состояния PP



— концептуализации, включающие в себя психические АКТы, могут служить причинами для других концептуализаций



— состояния или их изменения могут обеспечивать условия для концептуализаций



— один PP эквивалентен другому PP или является частным случаем другого

— АКТы могут варьироваться вдоль определенных измерений (напрмер, скорость для действий движения)

4.1.4. Концептуальные времена ТКЗ

В ТКЗ употребляется набор модификаторов концептуализаций, соответствующих временам в языке:

- ∅ — настоящее,
- *p* — прошедшее,
- *f* — будущее,
- / — отрицание,
- *ts* — начало существования,
- *tf* — конец существования,
- *c* — условное,
- *k* — продолжительное,
- ? — вопросительное,
- ∞ — постоянное.

4.1.5. Элементарные действия ТКЗ

Концептуальное действие – то, что может быть сделано некоторым деятелем над некоторым объектом. Различаются две категории действий: физические (над физическими объектами) и психические, или мыслительные (над идеями или идеальными сущностями, например, ощущениями - последние два из ниже приведенных).

PROPEL — *прикладывать силу к*, требует объекта (достаточно малого по отношению к силе) и директивного падежа, указывающего направление прикладываемой силы.

MOVE — *двигать частью тела*, требует директивного падежа для описания пути движения части тела.

INGEST — *принять что-то внутрь одушевленного объекта*, здесь объект должен быть меньше отверстия в теле деятеля.

EXPEL — *взять что-либо изнутри одушевленного объекта и вытащить наружу*, здесь объект должен быть предварительно принят внутрь.

GRAPS — *физически захватить объект*, здесь объект не должен превышать определенных размеров, директивный падеж указывает направление к той части тела, которая осуществляет захватывание.

PTRANS — *изменить местоположение чего-либо*, требует объектного, директивного и инструментального падежей.

ATRANS — *изменить некоторое абстрактное отношение для объекта*,

SPEAK — *произвести звук*, требует директивного падежа.

ATTEND — *направить орган чувств к определенному стимулу*, требует директивного падежа.

MTRANS — *передавать информацию*, здесь объекты — всегда концептуализации, требуется реципиентный падеж, где потенциальными получателями являются отделы человеческого мозга, а потенциальными донорами — органы чувств или отделы человеческого мозга.

MBUILD — *создавать или сочетать мысли*, здесь объекты — концептуализации, из которых (в результате MBUILD) создаются новые концептуализации.

4.1.6. Состояния объектов ТКЗ

Многие состояния в ТКЗ описываются посредством шкал, имеющих числовые значения. В качестве примера можно привести следующие шкалы.

ЗДОРОВЬЕ (HEALTH) — от −10 до +10:

- мертвый −10,
- смертельно больной −9,
- больной от −8 до −3,
- нездоровится −2,
- нормально 0,
- прекрасно +7,
- абсолютно здоров +10.

СТРАХ (FEAR) — от −10 до 0:

- в ужасе −9,
- напуган −5,
- встревожен −2,
- спокоен 0.

РАЗДРАЖЕНИЕ (ANGER) — от −10 до 0:

- рассвирепевший −9,
- разъярен −8,
- разгневан −6,
- раздражен −2,
- спокоен 0.

ПСИХИЧЕСКОЕ СОСТОЯНИЕ (MENTAL STATE) — от −10 до 10:

- в прострации −9,
- подавлен −5,
- расстроен −3,
- грустен −2,
- нормально 0,
- доволен +2,
- счастлив +8,
- в экстазе +10.

ФИЗИЧЕСКОЕ СОСТОЯНИЕ (PHYSICAL STATE) — от −10 до +10:

- мертв −10,

— сильные телесные повреждения −9,
— ранен −5,
— сломан (для объектов) −5,
— побит от −1 до −7,
— нормально +10.

СОЗНАНИЕ (CONSCIOUSNESS) — от 0 до +10:

— без сознания 0,
— сон +5,
— бодрствование +10.

ГОЛОД (HUNGER) — от −10 до +10:

— «умирающий от голода» −8,
— голоден как волк −6,
— голоден −3,
— нет аппетита 0,
— сыт +3,
— «сыт по горло» +6,
— «до отвала» +8.

ОТВРАЩЕНИЕ (DISGUST) — от −10 до 0:

— омерзительный −8,
— отвратительный −6,
— противный −4,
— надоевший −2.

УДИВЛЕНИЕ (SURPRISE) — от 0 до 10:

— удивлен +5,
— изумлен 7,
— потрясен +9.

Некоторое состояния не являются шкалами, а имеют обычные абсолютные меры. Таковыми являются, например, **ДЛИНА (LENGTH), ЦВЕТ (COLOR), ИНТЕНСИВНОСТЬ СВЕТА (LIGHT INTENSITY), МАССА (MASS), СКОРОСТЬ (SPEED)**.

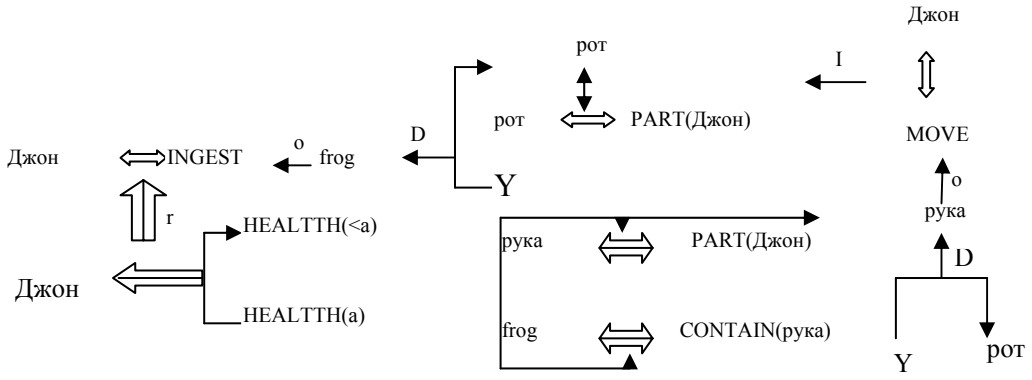
Существуют состояния, которые являются не шкалами, а просто отношениями между объектами. Например, **КОНТРОЛЬ (УПРАВЛЕНИЕ) (CONTROL), НЕОТЪЕМЛЕМОЕ ОБЛАДАНИЕ (ЧАСТЬ) (PART), ОБЛАДАНИЕ (POSS), СОБСТВЕННОСТЬ (OWNERSHIP), СОДЕРЖАТЬ (CONTAIN), БЛИЗОСТЬ (PROXIMITY), МЕСТОПОЛОЖЕНИЕ (LOCATION), ФИЗИЧЕСКИЙ КОНТАКТ (PHIS.CONT)** и др.

В ТКЗ используется еще одно отдельно стоящее состояние (**MFEEL**), которое выражает отношение между двумя людьми и эмоцией.

Отметим, что для использования (не только сравнения) состояний в процессе понимания текста в памяти системы нужно хранить детальное описание того, что означают все эти понятия.

4.1.7. Примеры работы системы MARGIE

Классический пример предложения, на котором Р. Шенк поясняет принципы построения смыслового представления – «John ate a frog» (Джон съел лягушку). В терминах ТКЗ смысл этого предложения выглядит так:



Пример работы программы в режиме умозаключения:

ВХОД: Джон сказал Мэри, что Биллу нужна книга.
ВЫХОД1: О чем книга?
ВЫХОД2: Мэри знает, что Биллу нужна книга.
ВЫХОД3: Билл хочет прийти за книгой.
ВЫХОД4: Билл хочет, чтобы некто перестал иметь книгу.
ВЫХОД5: Билл хочет прочитать книгу.

Пример работа программы в режиме перифразирования:

ВХОД: Джон посоветовал Мэри выпить вина.
ВЫХОД: Джон сказал Мэри, что ей станет лучше, если она выпьет вина.

4.2. Теория лингвистических моделей «СМЫСЛ⇔ТЕКСТ»

Теория лингвистических моделей «СМЫСЛ⇔ТЕКСТ» изложена в [13].

В соответствии с этой теорией ЕЯ рассматривается как особого рода преобразователь, выполняющий переработку заданных смыслов в соответствующие им тексты и заданных текстов в соответствующие им смыслы. Под смыслом понимается инвариант всех синонимичных преобразований (без доказательства его существования в общем случае), а синонимичным преобразованием называется переход от одного равнозначного текста (поставленному в соответствие одному и тому же явлению действительности) к другому.

Модели «СМЫСЛ⇔ТЕКСТ» — модели модульного типа, в них выделяются и отдельно описываются различные языковые уровни. На семантическом уровне исходной информацией является некоторое синтаксическое представление текста. Причем, в этих моделях различаются так называемый *глубинный* (семантизированный, учитывающий некоторые семантические отношения) синтаксис и *поверхностный*

(«чистый») синтаксис. Результатом же преобразований семантического уровня является определенное изображение содержания связного фрагмента речи без расчленения на фразы и словоформы — т.е. в виде семантического представления (которое и является записью смысла).

Семантическое представление состоит из двух компонентов: *семантического графа* (СГ) и сведений о *коммуникативной организации смысла* (КОС).

СГ представляет собой связанный ориентированный граф, вершины которого помечаются символами *сем*, а дуги изображают связи сем-предикатов с их аргументами. Стрелки направляются от предикатов к аргументам и нумеруются.

Семами называются элементарные смысловые единицы, атомы смысла, семантически различимые единицы. Различаются следующие типы сем:

— *кванторы (например, квантор существования)*,

— *логические связки (например, конъюнкция, отрицание)*,

— *имена предикатов или отношений (например, равенство)*,

— *предикатные переменные*,

— *имена объектов или классов*.

Одним из главных аспектов **КОС** является членение некоторой порции записи смысла на тему (то, о чем говорится) — **T** — и рему (то, что говорится) — **R**, а также определение различных логических акцентов.

Для работы семантического компонента, основанного на модели «СМЫСЛ⇔ТЕКСТ» необходимо создать **семантический язык** и **толково-комбинаторный словарь (ТКС)**. В этих моделях под семантическим языком понимается

а) *семантический словарь*, в который входит словарь элементарных семантических единиц — сем (имен предметов и предикатов), словарь промежуточных семантических единиц и словарь символов, характеризующих коммуникативную организацию смысла: тема — рема, старое — новое, выделено — не выделено и т.п.;

б) *правила образования*, по которым из семантического словаря могут строится семантические представления высказываний и которые касаются только формальной правильности семантических представлений;

в) *правила преобразования*, которые задают синонимичность двух семантических представлений.

Кроме того, для использования семантического языка необходимо иметь набор семантических аксиом и набор правил семантической «комбинаторики» — правил расчленения/сочленения семантических представлений при переходе от смысла к тексту и наоборот.

Словарная статья каждой **словарной единицы C_0** должна содержать все слова или словосочетания, определенным образом связанные с ней по смыслу, а именно:

1) ее «парадигматические варианты» или «замены» — языковые средства, которые могут или должны заменять C_0 в тех или иных контекстах и при тех или иных условиях;
2) ее «синтагматические партнеры» или «параметры» — языковые средства, которые идиоматично выражают при данной словарной единице некоторые смыслы.

Парадигматические варианты и синтагматические партнеры C_0 называются **лексическими коррелятами**.

Зависимости, связывающие слова с их лексическими коррелятами, в моделях «СМЫСЛ⇔ТЕКСТ» предлагается описывать с помощью лексических функций.

Лексическая функция (ЛФ) f описывает зависимость, определяющую для некоторого слова или словосочетания X такое множество слов или словосочетаний $\{Y_i\} = f(X)$, что для любых X_1 и X_2 верно следующее: если $f(X_1)$ и $f(X_2)$ существуют, то между $f(X_1)$ и X_1 с одной стороны, и между $f(X_2)$ и X_2 — с другой, всегда имеет место одно и то же смысловое отношение.

ЛФ вводятся как средство лексической сочетаемости, а не семантики.

Полный перечень лексических функций см. в [13], здесь же приводятся лишь некоторые из них в качестве примера.

Syn — синоним: слово, совпадающее с C_0 по смыслу, принадлежащее к той же части речи и имеющее такие же активные валентности;

Syn (лингвистика) = языкознание.

Conv — конверсив: слово, которое называет то же самое отношение, что и C_0, но взятое в «ином направлении», т.е. с перестановкой тех же актантов в другие места;

Conv (бояться) = пугать.

Anti — антоним: слово, обозначающее свойство, состояние или действие, «противоположное» свойству, состоянию или действию, обозначенному C_0;

Anti (друг) = враг, Anti (закрывать) = открывать.

Der — синтаксический дериват: слово, совпадающее с C_0 по смыслу, но принадлежащее к другой части речи;

Der (искать) = поиск.

Gener — название понятия, родового по отношению к понятию, обозначенному C_0;

Gener (клубника) = ягода.

Attr — «атрибут»;

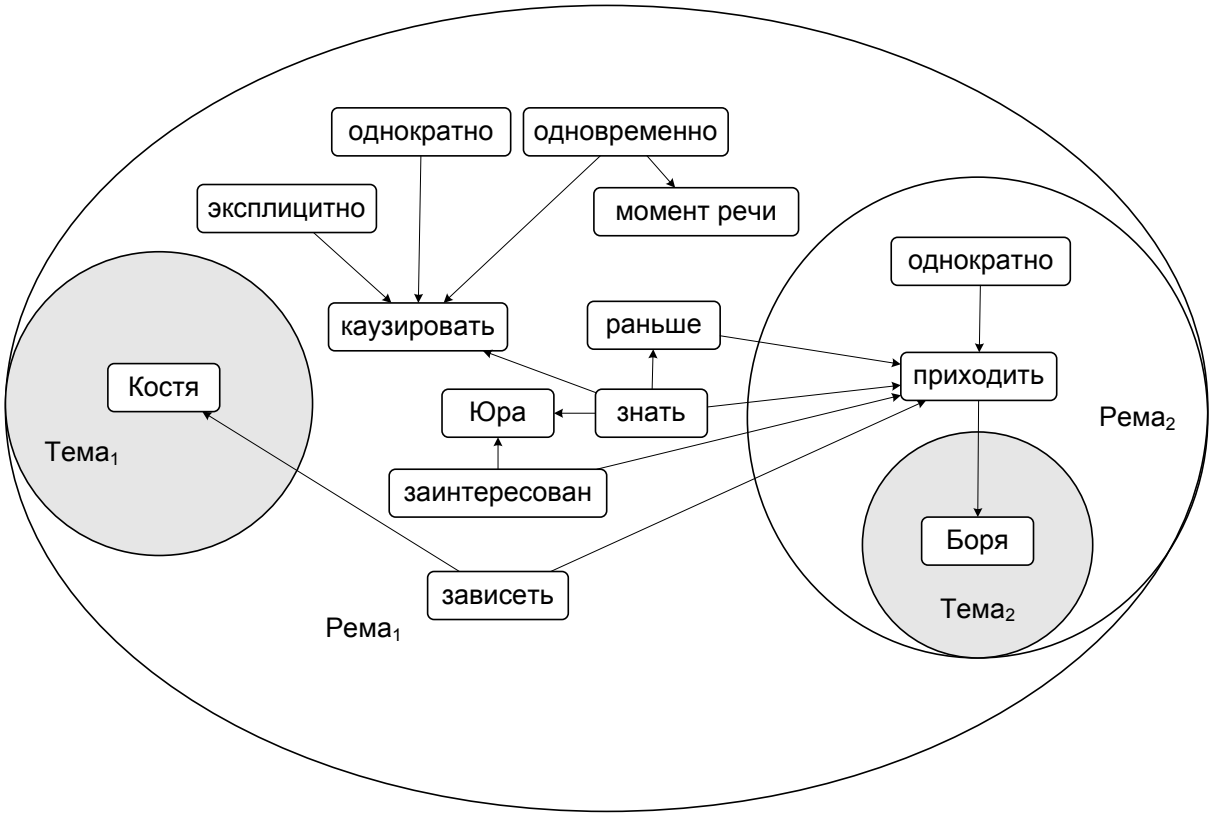
Attr (актер) = сцена.

Cous — «делать так, чтобы»;

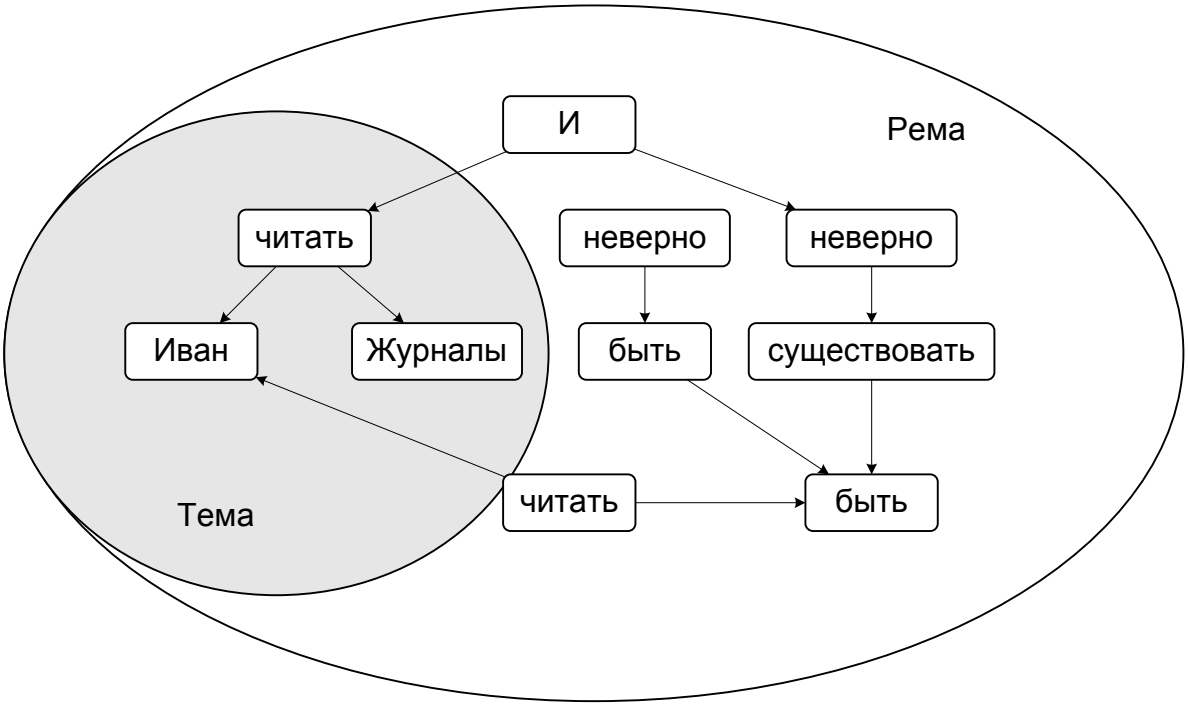
Cous (школа) = открывать.

Примеры семантических представлений:

1. *Костя обещает Юре, что Боря придет.*



2. *Иван читает только журналы.*



4.3. Падежные системы

Среди АОТ-систем, имеющих семантический компонент выделяются так называемые падежные системы, в которых семантика предложения описывается как система семантических валентностей, через связи главного слова с его семантическими актантами, т.е. через семантические отношения или семантические падежи. Например,

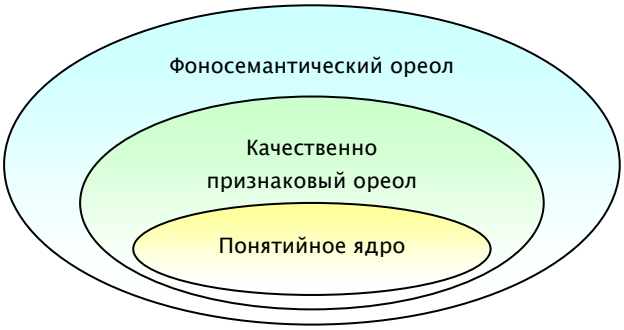
глагол *«дать»* требует семантических падежей *дающего* (агенса), *адресата* и *объекта передачи*.

Слабым местом падежных систем является то, что в них, как правило, отсутствуют четкие определения и критерия выделения падежей, неясна степень полноты их набора и границы между «падежными» и другими элементами предложения.

4.4. Звук и смысл

Вообще говоря, чтобы правильно использовать слова в разных ситуациях, недостаточно знать значения слов, их словарные толкования. Так, например, обозначающие одни и те же понятия слова *«отец»* и *«папа»*, *«бабушка»* и *«старуха»* в разных ситуациях далеко не взаимозаменимы. **Понятийное ядро** слов как бы окружено ореолами созначений. Ближайший к ядру ореол — **качественно-признаковое значение** слова. Это тот аспект значения, который можно описать путем перечисления качественных признаков данного понятия. Например, понятийное ядро слова *«птица»* — покрытое пухом и перьями животное из класса позвоночных с крыльями, двумя ногами и клювом, а качественно-признаковое значение — что-то быстрое, стремительное. У других понятий на первое место могут выдвигаться другие признаки, например, *«слон»* — большой и сильный, *«любовь»* — хорошее, светлое, возвышенное. Зная качественно-признаковое значение слов, можно понять, например, фразу *«Он птицей полетел на свидание»* или *«Что за прелесть!»*, которая может относиться и к девушке, и к платью, и к закату, и к цветку…

Кроме того, и звуковая форма слов содержательна, информативна, служит одним из аспектов значения. Звуковая содержательность чаще всего поддерживает, подчеркивает понятийный и качественно-признаковый аспекты значения. Соответствие значения и звучания делает слово выразительнее, живее, «нагляднее». Ореол содержательности звучания, или фоносемантический ореол, есть у любого слова, только чаще всего мы его не замечаем.



Вообще говоря, в языке информативно, пропитано значением все — звучание речи, отдельные слова и их сочетания, способы соединения слов, устройство предложений, композиция текстов — все от мельчайших звуков до сложной архитектуры всего здания языка.

Оказалось, что самые тонкие и трудно уловимые ореолы значения поддаются числовому измерению, а значит, компьютерной обработке.

Около 40 лет назад группа американских исследователей под руководством Ч. Осгуда опубликовала сенсационную книгу под заглавием «Измерение значения». Осгуд доказал, что в области семантики возможны измерения, и показал, как их можно выполнить. С помощью методики Осгуда оказалось возможным измерить качественный аспект (ореол) значения. Свой измерительный инструмент Осгуд назвал «семантический дифференциал» — это некоторая шкала, например «хорошее — плохое»:

1 – очень хорошее

2 – хорошее

3 – никакое

4 – плохое

5 – очень плохое.

Шкала предлагается носителям языка — информантам (не менее 50-60 человек), которым предъявляются слова, и они должны поставить словам оценки. По ответам информантов вычисляется средняя оценка слова по данной шкале. Например, оценка слова «дом» по шкале «хорошее — плохое» — 2.2.

Чтобы получить полное описание качественно-признаковой ореольной семантики языка, надо брать шкалы из всех прилагательных и мерить по ним все остальные слова языка. Это огромная работа. К счастью, оказалось, что есть всего три основные шкалы, три фактора, вокруг которых группируются все остальные (т.е. по ним получаются почти такие же значения). Это

1. Фактор оценки (хорошее плохое, полезное — вредное, светлое — темное, красивое — безобразное…),

2. Фактор силы (сильное — слабое, легкое — тяжелое, большое — маленькое…),

3. Фактор активности (подвижное — статичное, быстрое — медленное, активное — пассивное…).

А в русском языке выделяется еще один фактор:

4. Фактор родокомфортности (нежное — грубое, мягкое — твердое, женственное — мужественное, удобное — неудобное…).

Т.е. получается, что качественно-признаковому ореолу каждого слова соответствует точка в трехмерном (четырехмерном) кубе, и координаты точки можно измерить. Но работа по составлению такого куба достаточно трудоемка и сопряжена с рядом сложностей.

Выяснился поразительный факт, что «семантическим дифференциалом» можно «измерить» и звуки речи. Исследованием проблемы соотношения звука и смысла в России занимается группа ученых под руководством профессора А.П. Журавлева [14].

При «измерении» звуков информанты ставили им оценки по тем же признаковым шкалам, которые объединились в те же группы (факторы). В результате выявились непреложные закономерности: О — светлый, Ы — темный, Р — устрашающий, И — безопасный, К — быстрый, Ш — медленный… После измерения всех звуков по четырем шкалам и размещения их в семантическом кубе получилось русское фонетико-семантическое пространство. Имея его компьютер моментально вычисляет фоносемантический ореол любого слова (учитывая частоту встречаемости каждого звука в речи, ударение и первую букву слова).

Конечно же, нет прямой связи между понятиями о предметах, предметами и их названиями. А вот качественный и фоносемантический ореолы, как правило, совпадают. Т.е. у большинства слов значение и звучание находятся во взаимном соответствии, и опыт показывает, что дисгармония аспектов значения затрудняет функционирование слов в речи.

Вот несколько наиболее ярких примеров вычисления фоносемантического ореола слов:

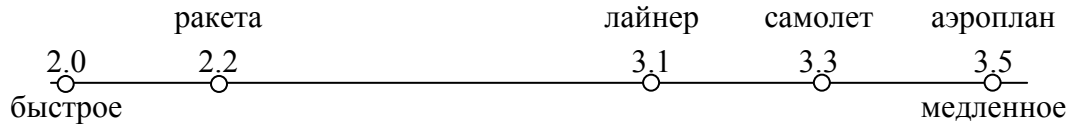
БОГАТЫРЬ — большое, мужественное, сильное, громкое, могучее.

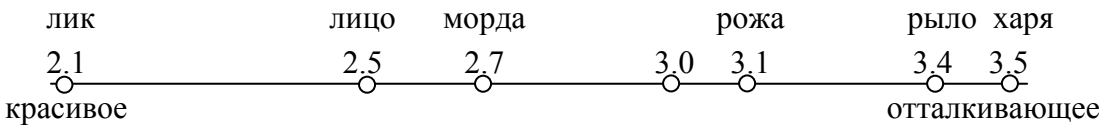
ИДЕАЛ — хорошее, светлое, красивое, яркое, доброе.

ЛАСКА — хорошее, гладкое.

ШУШЕРА — плохое, темное, отталкивающее, низменное, тусклое.

А вот интересное соотношение фоносемантических оценок близких по смыслу слов по некоторым показательным шкалам:





5. Прагматический компонент ЕЯ-систем

Прагматика — наука, изучающая язык в его отношении к тем, кто его использует, функционирование языковых знаков в речи.

Пропозиция — факт, передаваемый предложением; смысловое содержание предложения; семантический инвариант, общий для всех членов модальной и коммуникативной парадигм предложения и производных от предложения конструкций.

Пропозиция является **пресуппозицией** в прагматическом смысле, если говорящий считает ее истинность само собой разумеющейся и исходит из того, что другие участники контекста считают также.

Набор пресуппозиций человека определяется на основании тех утверждений, которые он делает, вопросов, которые он задает. Пресуппозиции — это пропозиции, неявно подразумеваемые еще до начала передачи речевой информации. Пресуппозиции, естественно, не обязаны быть истинными. В нормальной ситуации человек, по меньшей мере, убежден в истинности своих пресуппозиций. Именно поэтому мы часто извлекаем из высказываний данного лица больше информации о его убеждениях, чем он нам сообщает. Иногда, однако, в пресуппозиции может быть пропозиция, в истинности которой мы сомневаемся, или даже про которую считаем или знаем, что она ложна. Это происходит в ситуации обмана.

Модальность — отношение значения предложения к действительности.

Вербализация — речевое выражение чего-либо.

Анафора (anaphora (греч.)) — повторение.

Анафорический — отсылочный.

Номинация (лат. nominatio — наименование) — образование языковых единиц, характеризующихся номинативной функцией, т.е. служащих для названия и вычленения фрагментов действительности и формирования соответствующих понятий о них в форме слов, сочетаний слов, фразеологизмов и предложений.

Номинат — представление говорящего о фрагменте ситуации, для которой производится вербализация.

5.1. Предмет изучения прагматического компонента

Перед прагматическим компонентом встает два рода проблем:

- определение типов речевых актов, «продуктов» речи и
- описание признаков и свойств речевого контекста, влияющих на определение того, какая именно пропозиция выражается данным предложением.

Различаются следующие основные типы речевых актов:

- утверждение,
- приказ,
- контрфактическое высказывание,

— требование,
— догадка,
— опровержение,
— просьба,
— возражение,
— предсказание,
— обещание,
— призыв,
— рассуждение,
— объяснение,
— оскорбление,
— вывод,
— умозаключение,
— предположение,
— обобщение,
— ответ,
— обман.

Задача определения типа речевого акта считается задачей, не представляющей особых трудностей, если известен контекст высказывания, влияющий на ту функцию, которую говорящий придает определенной пропозиции, а также на саму пропозицию. Очевидно, что семантические правила определения пропозиции, выражаемой данным предложением, могут быть сформулированы лишь с учетом некоторых признаков той ситуации, в которой это предложение произносится. В качестве примеров можно проанализировать следующие предложения:

«Все хорошо проводят время».

«Мы победим».

«Это великолепная картина».

«Ты можешь передать мне солонку?».

Задача определения типа речевого акта в итоге сводится к заданию необходимых и достаточных условий успешного выполнения целей самого речевого акта, а не определению условий истинности, заключенной в нем пропозиции.

Задача описания признаков и свойств речевого контекста сводится к изучению того, каким образом речевой контекст влияет на пропозицию, выражаемую предложением в этом контексте.

В целом лингвистическая прагматика не имеет четких контуров, в нее включается комплекс вопросов, связанных с говорящим субъектом, адресатом, их взаимодействием в коммуникации, ситуацией общения.

В связи с субъектом речи изучаются:

1) явные и скрытые цели высказывания,
2) речевая тактика и типы речевого поведения,
3) правила разговора, подчиненные так называемому принципу сотрудничества, рекомендующему строить речевое общение в соответствии с принятой целью и направлением разговора, например:
    а) адекватно нормировать сообщаемую информацию,
    б) сообщать только истинную информацию и обоснованные оценки,
    в) делать сообщения релевантными относительно темы разговора,

г) делать речь ясной, недвусмысленной и последовательной,

4) установка говорящего или прагматическое значение высказывания: косвенные смыслы высказывания, намеки, иносказание, обиняки и т.п.,
5) референция говорящего, т.е. отнесение языковых выражений к предметам действительности, вытекающие из намерения говорящего,
6) прагматические пресуппозиции: оценка говорящим общего фонда знаний, конкретной информированности, интересов, мнений и взглядов, психологического состояния, особенностей характера и способности понимания адресата,
7) отношение говорящего к тому, что он сообщает:
    а) оценка содержания высказывания (его истинность, ирония, многозначительность, несерьезность и пр.),
    б) введение в фокус интереса одного из тех лиц, о которых говорящий ведет речь (или эмпатия),
    в) организация высказывания в соответствии с тем, чему в сообщении придается наибольшее значение.

В связи с адресатом речи изучаются:

1) интерпретация речи, в том числе правила вывода косвенных и скрытых смыслов из прямого значения высказывания,
2) воздействие высказывания на адресата: расширение информированности адресата, изменения в его эмоциональном состоянии, взглядах и оценках, влияния на совершаемые адресатом действия, эстетический эффект и т.п.,
3) типы речевого реагирования на высказывание (например, способы уклонения от прямого ответа на вопрос).

В связи с отношениями между участниками коммуникации изучаются:

1) формы речевого общения (информативный диалог, дружеская беседа, спор, ссора и т.п.),
2) социально-этикетная сторона речи (формы обращения, стиль общения),
3) соотношение между участниками коммуникации в тех или иных речевых актах (сравните просьбу и приказ).

В связи с ситуацией общения изучаются:

1) интерпретация таких слов, как *«здесь»*, *«сейчас»*, *«этот»* и т.п.
2) влияние речевой ситуации на тематику и формы коммуникации (типичные темы и формы разговоров, например, в гостях, в больницах, в приемных адвокатов и т.п.).

Изучение прагматических свойств языка имеет обширные области пересечения с изучением свойств связного текста.

5.2. Анализ связного текста (дискурса)

Текст — комплекс взаимосвязанных друг с другом предложений, который обладает определенной автономностью от прочих аналогичных комплексов и определенной смысловой целостностью, единством, обусловливаемым единством коммуникативного намерения автора текста.

Целостность текста превращает его в систему, в которой элементы зависят друг от друга и предполагают друг друга, что проявляется в двух типах явлений: в пронизывающих и скрепляющих текст повторениях некоторых смысловых единиц и в опущении отдельных фрагментов текста, которые в изолированном предложении были бы необходимы, а в целом тексте могут быть восполнены из других предложений.

Наиболее часто разработчики автоматизированных систем сталкиваются с явлениями *анафорического* (отсылочного, *анафора* — повтор) *замещения* значащих выражений местоимениями и *эллипсиса* (опущения структурно необходимых частей предложения, когда они аналогичны каким-то частям данного или соседнего предложений). Но это лишь часть широкого круга явлений сверхфразового уровня. Так, проблема анализа анафорических связей местоимений — частный случай проблемы выявления анафорических отношений в тексте и связанных с ней проблем **референции** — установления соответствия между языковыми выражениями (прежде всего именными группами) и внеязыковыми сущностями, и **кореференции** — определения языковых выражений, эквивалентных с точки зрения референции.

Дескрипция — именная группа, в вершине которой стоит нарицательное существительное.

Полная дескрипция — наиболее простой способ осуществления однозначной референции к некоторому объекту. Например, *«Будьте ровно в восемь часов около канавы, в которую вчера упала с головы вашей шляпа»*.

Неполным дескрипциям заведомо удовлетворяют более одного объекта. В случае референции с помощью неполных дескрипций или с помощью неоднозначной анафорической номинации говорят, что имеет место референциальный конфликт. Например, *«Рассмотрим структуру памяти ЭВМ, которая состоит их двух основных частей»*.

Одной из основных моделей обработки дискурса, на которой основываются современные системы обработки связного текста, является модель Т.А. ван Дейка и В. Кинча [15, 16]. Основные положения этой модели следующие.

Сутью процессов обработки дискурса является конструирование в памяти человека ментального представления текста. В этих процессах используются как воспринимаемая информация, так и информация, содержащаяся в памяти человека. Люди способны сконструировать ментальное представление текста, описывающего объекты, события, при условии, что они имеют более общие знания о таких объектах и событиях.

Существуют различные типы связного текста. Каждый тип текста имеет свои языковые и когнитивные (познавательные) различия. Существуют различные типы, стили и способы понимания (например, быстрый просмотр газеты, чтение учебника). Разные пользователи языка могут обладать различными знаниями и умениями, что влияет на понимание дискурса. В разных ситуациях пользователи языка по-разному понимают разные тексты.

Важной особенностью понимания связных текстов является способность человека предугадывать их содержание. У понимающего текст возникают ожидания того, что будет сказано, прежде чем он реально это услышит или прочитает, и это может облегчить процесс понимания.

Один из принципиальных моментов модели ван Дейка и Кинча — выделение и обособление текстовых баз и ситуационных моделей. С одной стороны, в процессе понимания текста строится его текстовая база, которая включает семантическое представление воспринимаемого текста, и в виде взаимосвязанных пропозиций фиксирует то, о чем было сказано в тексте. Текстовая база фиксирует и особенности формы выражения пропозиций на ЕЯ.

С другой стороны, понимание текста предполагает активацию, обновление и другие формы функционирования так называемой ситуационной модели: это когнитивное представление событий, действий, лиц и вообще ситуаций, о которых говорится в тексте. Текстовая база сопоставляется с тем, что есть в ситуационной модели, а

ситуационная модель может модифицироваться в соответствии с содержанием текстовой базы.

Отметим, что формируемая в процессе понимания текстовая база должна удовлетворять критерию *локальной связности*, предполагающему, что пропозиции, встречающиеся в тексте должны быть связаны между собой, а также критерию *глобальной связности*. В процессе понимания дискурса человек пытается понять не только локальную информацию, но и текст в целом, выявить его суть, содержание, тему. Глобальная связность предполагает, что в процессе понимания выявляется основная тема текста и определяется роль каждой фразы с точки зрения раскрытия основной темы.

Еще одним важным фактором, влияющим на процессы понимания дискурса, является учет *схематической структуры* текста. Во многих типах дискурса проявляется традиционная, культурно-обусловленная схематическая структура, организующая глобальное содержание текста. Например, рассказы, как правило, строятся по иерархической схеме *Завязка – Кульминация – Развязка*. Свою структуру имеют доказательства чего-либо, описания сложных технических систем и прочее.

ЛИТЕРАТУРА

[1]. Зализняк А.А. Грамматический словарь русского языка. — М.: Русские словари, 2003.

[2]. Волкова И.А. Адаптация и обучение системы общения с ЭВМ на естественном языке. Канд. диссертация. М. 1982. — http://axofiber.org.ru/download/volkova-dissertation.pdf

[3]. Попов Э.В. Обыщение с ЭВМ на естественном языке. — М.: Наука, 1982.

[4]. Искусственный интеллект. Книга 1. Справочник. — М.: Радио и связь, 1990.

[5]. Мальковский М.Г. Диалог с системой искусственного интеллекта. — М.: Изд-во МГУ, 1985.

[6]. Апресян Ю.Д. Идеи и методы современной структурной лингвистики. — М.: Просвещение, 1966.

[7]. Гладкий А.В. Синтаксические структуры естественного языка в автоматизированных системах общения. — М.: Наука, 1985.

[8]. Шенк З. Обработка концептуальной информации. — М.: Энергия, 1980.

[9]. Лингвистический энциклопедический словарь. — М.: Советская энциклопедия, 1990.

[10]. Волкова И.А., Руденко Т.В. Формальные грамматики и языки. Элементы теории трансляции. — М.: Изд-во МГУ, 1999.

[11]. Ахо А., Сети Р., Ульман Дж. Компиляторы — М.: С-П, Киев, Вильямс, 2001.

[12]. Волкова И.А., Головин И.Г. Синтаксический анализ фраз естественного языка на основе сетевой грамматики. ДИАЛОГ'98, Труды межд. семинара. — М.: 1998.

[13]. Мельчук И.А. Опыт теории лингвистических моделей «СМЫСЛ ↔ ТЕКСТ». — М.: Наука, 1974.

[14]. Журавлев А.П. Звук и смысл. — М.: Просвещение, 1991.

[15]. Ван Дейк Т.А., Кинч В. Стратегия понимания связного текста.// Новое в зарубежной лингвистике. Вып. XXIII — М.: Прогресс, 1988, с. 153-211.

[16]. Ван Дейк Т.А. Язык. Познание. Коммуникация. — М.: 1989.