КОМПЬЮТЕРНЫЙ АНАЛИЗ ЕСТЕСТВЕННО-ЯЗЫКОВОГО ТЕКСТА

Рубашкин Валерий Шлемович,
д. техн. н., профессор

Митрофанова Ольга Александровна,
канд. филол. н., доцент

Литература

1. Palmer F. R. Semantics. A new outline. M., 1982.
2. Кобозева И. М. Лингвистическая семантика. М., 2000.
3. Кронгауз М. А. Семантика. М., 2001.
4. Лайонз Дж. Лингвистическая семантика: Введение. М., 2003.
5. Рубашкин В. Ш. Представление и анализ смысла в интеллектуальных информационных системах. М., 1989.
6. Nirenburg S., Raskin V. Ontological Semantics. – Cambridge, MA: MIT Press, 2004
7. Тузов В. А. Компьютерная семантика русского языка.- СПб.: Изд-во СПбГУ, 2003.
8. Леонтьева Н.Н. Автоматическое понимание текстов: системы, модели, ресурсы. – М: Издательский центр «Академия», 2006
9. Agirre E., Edmonds Ph. (eds). Word Sense Disambiguation. Algorithms and Applications - Springer, 2006.

0. Рубашкин В. Ш. Семантический компонент в системах понимания текста // КИИ-2006. Десятая национальная конференция по искусственному интеллекту с международным участием. Труды конференции. – М.: Физматлит, 2006

1. Рубашкин В. Ш. Словарная поддержка процедур семантической интерпретации предложных связей // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции "Диалог'2005". М., 2005. С. 430 – 435.

2. Рубашкин В. Ш. Универсальный понятийный словарь: функциональность и средства ведения // КИИ-2002. Восьмая национальная конференция по искусственному интеллекту с международным участием. Труды конференции. М., 2002. С. 231 – 237.

3. Рубашкин В. Ш., Чуприн Б.Ю. Распознавание количественной информации в ЕЯ-текстах // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции "Диалог 2006". – М.: Изд-во РГГУ, 2006. С. 456 – 458.
4. Рубашкин В. Ш. Прикладная лингвистика и языковая инженерия. // Труды международной конференции «Megaling'2005. Прикладная лингвистика в поисках новых путей». – СПб: Издательство "Осипов", 2005. С 115 – 123.
5. Виды неоднозначностей в размеченных корпусах и методы их разрешения // Труды международной конференции "Корпусная лингвистика-2006". – СПб.: Изд-во С.-Петерб. Ун-та, 2006, – С. 339 – 346.

Дополнительная литература

1. Арутюнова Н. Д. Предложение и его смысл (логико-семантические проблемы). М., 2003.
2. Гершензон Л. М., Ножов И. М., Панкратов Д. В. Система извлечения и поиска структурированной информации из больших текстовых массивов СМИ. Архитектурные и лингвистические особенности // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции "Диалог'2005". М., 2005. С. 97 – 101.
3. Ермаков А. Е. Референция обозначения персон и организаций в русскоязычных текстах СМИ: эмпирические закономерности для компьютерного анализа.// Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции "Диалог'2005". М., 2005. С. 131 – 135.

4. Кузнецов И. П. Методы обработки сводок с выделением особенностей фигурантов и происшествий // Труды международного семинара "Диалог-1999" по компьютерной лингвистике и ее приложениям. Т. 2. М., 1999.
5. Лебедев М.В., Черняк А. З. Онтологические проблемы референции. М., 2001.
6. Падучева Е. В. Высказывание и его соотнесенность с действительностью. М., 2004.
7. Рахилина Е. В. Когнитивный анализ предметных имен: семантика и сочетаемость. М., 2000.
8. Information Extraction. (Электронные документы)

Раздел 1. ВВЕДЕНИЕ В ДИСЦИПЛИНУ

Тема 1. Методологические основания

1. Автоматический анализ текста как инженерная задача.

Результат – программная система (=инженерная конструкция)

Прикладная лингвистика и общая лингвистика ***vs*** языковая инженерия (пересечение понятий)

Инженерия вообще – "техника" ***vs*** "естествознание":

объектные ***знания-что*** ***vs*** процедурные ***знания-как***

Знания-что: объекты, их свойства, отношения, процессы:

Где находится остров Тасмания?

Из чего состоит атом?

Знания-как: методы, способы, средства, инструменты:

Как сварить украинский борщ?

Как быстро вылечить ожог?

Что такое фотолитография?

Инженерная задача не имеет дисциплинарной принадлежности!

Общеизвестные примеры.

- Водный транспорт (судостроение): дерево – сталь; плотник – слесарь (клепка) – сварщик – наладчик сварочных автоматов.
- Воздушный транспорт (строительство летательных аппаратов): воздухоплавание (легкие газы, газонепроницаемые оболочки) - летательные аппараты, использующие подъемную силу крыла – вертолеты. Физика газов – аэродинамика; прочность и технология тканей и пленок – прочность и технология легких металлов. Винтовая и реактивная авиация

Автоматический анализ текста и вообще естественноязыковый диалог "человек – компьютер" не самоцель, а "неизбежное зло".

Следует избегать всюду, где это возможно, заменяя **регламентированным диалогом.**

Примеры:

- Общение с Word'ом: "диалоговое окно"
- Билетная справка и др. справочные системы.
- Медицинская диагностика
- Системы управления производством, банковские системы и др.
- Даже (предположительно) интерактивная энциклопедия – возможность движения от общего к частному.

О терминологии (и не только…)

Избыток названий, именующих разные разделы и направления, с одной стороны, и отсутствие единого их понимания, с другой:

- *прикладная лингвистика,*
- *структурная лингвистика,*
- *математическая лингвистика,*
- *компьютерная лингвистика,*
- *инженерная лингвистика,*
- *онтологическая семантика,*
- *корпусная лингвистика,*

наконец,

- *теоретическая лингвистика и общая лингвистика (ОТИПЛ)…*

– это действительно о разном?

Дисциплинарное окружение "прикладной" лингвистики –
та же картина:

- *искусственный интеллект,*
- *инженерия знаний,*
- *концептуальное моделирование,*
- *формальные (вычислительные) онтологии,*
- *философская логика,*
- *логическая семантика*
- *информационные технологии*

Ключевые противопоставления:

- *общая* ***vs*** *прикладная лингвистика;*
- *компьютерная* ***vs*** *"бескомпьютерная" лингвистика;*
- *структурная* ***vs*** *прецедентная (статистическая) лингвистика;*
- *лингвистическая* ***vs*** *"экстралингвистическая" ("концептуальная"?, "онтологическая") семантика.*

Общая и прикладная лингвистика

А.Н. Баранов:

прикладная лингвистика как "деятельность по приложению научных знаний об устройстве и функционировании языка *в нелингвистических научных дисциплинах* и в различных сферах практической деятельности человека, а также теоретическое осмысление такой деятельности".

Общая лингвистика - *знания-что* (как устроен и функционирует язык)

Прикладная лингвистика - *знания-как* (как эффективно учить языку; как переводить; как составлять словари; как моделировать на компьютере разные аспекты языковой компетенции человека)

Что касается *применения в нелингвистических научных дисциплинах* – ср., например, *физическую химию* (применение теоретических моделей и экспериментальных методов физики в химии).

Ср. также *психолингвистика, социолингвистика* и др.

Компьютерная - "бескомпьютерная" лингвистика.

Термин *компьютерная лингвистика* - если понимать его в прямом значении – в сегодняшней ситуации скорее дезориентирует, чем что-либо проясняет; он себя изжил.

Определения *прикладная, структурная, математическая, компьютерная* призваны были в 50-х – 60-х - 70-х г.г. прошлого века обозначить переход на новый уровень лингвистических исследований. Они – эти определения – были нужны, пока новые методы и подходы должны были отстаивать свое право на существование и как-то обозначать свою новизну и специфичность.

Фактически термин *компьютерная лингвистика* имеет в виду не просто *лингвистическое исследование с использованием компьютера*, а ***инженерное (с помощью компьютерных программ) моделирование разных аспектов языковой компетенции***. А для **этого** содержания более адекватным будет, термин ***инженерная лингвистика.***

Инженерная лингвистика, по-видимому, не теряя связи с общей лингвистикой, все более будет смыкаться с инженерией знаний, особенно на семантическом уровне.

Строго говоря, инженерная лингвистика –

это не совсем лингвистика, или, точнее, **не только лингвистика**.

Термины *прикладная* и *инженерная лингвистика* должны быть соотнесены не как общее и частное, а как

два понятия с пересекающимися объемами.

"Математический лингвист – это человек, который применяет то немногое, что он знает из математики к тому немногому, что он знает из лингвистики" (*конец 1950-х ?)*

Резюме – достаточно 3-х терминов:

Общая лингвистика, прикладная лингвистика, ***языковая инженерия*** (условно - *инженерная лингвистика*).

Еще один термин:

ICSC2007
First IEEE International Conference on **Semantic Computing**
September 17-19, 2007
Irvine, California, USA
http://ICSC2007.eecs.uci.edu

The field **Semantic Computing** applies technologies in natural language processing, data and knowledge engineering, software engineering, computer systems and networks, signal processing and pattern recognition, and any combination of the above to
extract, access, transform and synthesize the semantics (contents) of multimedia, texts, services and structured data.

Topics for submission include but are not limited to:

Natural language understanding and processing
Understanding and processing of texts and multimedia contents
Content-based retrieval of texts, images, videos and audios
Speech recognition
Semantic web search and services
Semantic services engineering
Semantic annotation of multimedia contents
Natural language driven computing
Multimedia driven computing
Question answering
Spoken dialogue and multi-modal systems
Data, knowledge and software engineering issues
Integration of semantic systems
Semantic computing and wireless communications
Content-based security
Applications of semantic computing
Hardware support for semantic computing systems

Тема 2. Проблемы и ограничения. Реальные задачи семантического анализа

Начало XXI века (2010-е и 2020-е) – эпоха лингвистических информационных технологий!

2.1. Реальные задачи семантического анализа

Общая цель семантического анализа – обеспечить понимание любого осмысленного текста.

Операциональная конкретизация: переход от плохо структурированной (ЕЯ-текст) к хорошо структурированной информации, пригодной для обработки стандартными и высокоэффективными средствами информационных технологий.

а) Общие задачи - дополнительная поддержка большинства лингвистических ИТ

Основные лингвистические технологии:

- Автоматический перевод – первая "лингвистическая" информационная технология.
- Документальные информационные системы.
- Технологии распознавания письменных текстов и устной речи.
- Орфографические и грамматические корректоры.
- Системы понимания (смыслового анализа и синтеза) текста.

Общие задачи:

– дополнительные лингвистические фильтры (в системах распознавания - OCR и Speech Recognition; в корректорах)

– разрешение неоднозначностей (в системах перевода и др.)

– дополнительные критерии релевантности документа (в документальных ИПС)

b) Специфическая задача:

Переход от плохо структурированной (ЕЯ-текст) к хорошо структурированной информации.

Целевые технологии:

- СУБД (формализация фактологической информации)
- Экспертные системы и онтологии (формализация номологической информации)
- В перспективе – перевод с профессионального языка на логический язык (куда специализированные ЯПЗ должны быть интегрированы) - с использованием машины ограниченного вывода.

Типовая задача **сегодняшнего дня:**

извлечение из ЕЯ-текстов фактографической информации и структурирование ее, например, в форме записей РБД, XML-разметки и т.п.
(***Information Extraction / Text Mining***).

Объект анализа - ситуативные ("планшетные") тексты:

- сообщения о движении и грузообработке судов;
- сообщения о криминальных происшествиях;
- медицинская карта;
- сообщения о расположении и состоянии сил и средств, участвующих в военных действиях;
- мониторинг общественно-политической / финансово экономической ситуации;
- рекламные сообщения и т. п.
- молекулярная биология: экспрессия генов.

Jerry R. Hobbs, Douglas Appelt, John Bear,
David Israel, Megumi Kameyama, Mark Stickel, and Mabry Tyson

Artificial Intelligence Center
SRI International
Menlo Park, California

FASTUS is a system for extracting information from natural language text for
entry into a database and for other applications. It works essentially as a cascaded, nondeterministic finite-state automaton.

There are five stages in the operation of FASTUS.

Stage 1: Names and other fixed form expressions are recognized.

Stage 2: Basic noun groups, verb groups, and prepositions and some other particles are recognized.

Stage 3: Certain complex noun groups and verb groups are constructed.

Stage 4: Patterns for events of interest are identified in and corresponding ``event structures'' are built.

Stage 5: Distinct event structures that describe the same event are identified and merged, and these are used in generating database entries.

This decomposition of language processing enables the system to do exactly the right amount of domain-independent syntax,
so that domain-dependent semantic and pragmatic processing
can be applied to the right larger-scale structures.

FASTUS is very efficient and effective, and has been used successfully in a number of applications.

Другая типовая задача -

формализация нормативных документов разного типа –

в частности, нормативно-технической (СНИПы, ГОСТы…)

и юридической документации.

Цель формализации, например, - проверка непротиворечивости корпуса нормативных актов; проверка логического соответствия вновь принимаемого нормативного акта существующей нормативной базе.

Пример постановки задачи типа *Information Extraction*:

Распознаваемые факторы:

3 Уровень налогов в Латвии
10 Число пенсионеров в Латвии
14 Объем экспорта Латвии на рынки ЕС
20 Уровень инфляции в Латвии (%)
23 Средняя заработная плата в Латвии
34 Уровень безработицы в Латвии (%)
55 Доступность образования в Латвии
56 Уровень подготовки специалистов в Латвии
72 Средний уровень пенсий в Латвии
80 Финансирование Латвии Евросоюзом
87 Уровень давления ЕС на Латвию (по вопросу о гражданских правах нацменьшинств)
100 Активность неграждан по защите своих прав и свобод

Релевантные контексты для фактора 100

Активность неграждан по защите своих прав и свобод

1001181

На минувшей неделе в Риге прошла ***забастовка русскоязычных школьников***

1001182

В начале марта в Риге пройдет Вселатвийский ***съезд защитников русских школ****.*

1001371

Волна протеста против ассимиляционной реформы *достигла своего апогея.*

1001714

После съезда наконец ***будет создана партия****, реально* ***защищающая интересы русских*** *Латвии.*

Релевантные контексты для фактора 80

Финансирование стран Балтии Евросоюзом

8001101

*За первые три года Латвия рассчитывает получить из общего бюджета ЕС **1,116 млрд. латов**.*

8001107

*В 2001-2002 гг. литовский сейм уже ратифицировал два договора с ЕС, благодаря которым в рамках программы SAPARD на развитие сельского хозяйства Литва получила **277,1 млн. литов**.*

1001371

*Со вступлением Эстонии в Европейский союз восточная граница страны станет одновременно и внешней границей ЕС. В связи с этим в 2004-2006 году ЕС планирует выделить из своего бюджета на финансирование укрепления восточной границы **около миллиарда эстонских крон**.*

Пример формализации технической нормы:

Жилые комнаты общежитий ***следует проектировать*** из расчета заселения не более трех человек при площади не менее 6,0 кв. м. на каждого проживающего. Комнаты ***должны быть*** непроходными, шириной не менее 2.2 м., их ***следует оборудовать*** встроенными шкафами площадью не менее 0.5 кв. м. на каждого проживающего.

(СНИП «Жилые здания»)

Общая структура нормы:

1) Нормируемый объект: *жилые комнаты общежитий*
2) Модальность предписания (*должны быть - допускается*)
3) Нормируемая характеристика:
4) Значение нормируемой характеристики

Нормируемые характеристики:

- *расчетная вместимость*: (*не более трех человек*);
- *площадь на проживающего:* *(не менее 6,0 кв. м);*
- *ширина:* *(не менее 2.2 м);*
- *проходная?*:* *(НЕТ);*
- *площадь встроенных шкафов на каждого проживающего:* *(не менее 0.5 кв. м.);*

Возможные запросы:

2. Нормируется ли указанный в запросе объект? – с учетом отношений род – вид.)
3. Какие объекты нормируются по данной характеристике?
4. Каковы допустимые значения указанной характеристики для указанного объекта?

И т. п.

2.2. Существенные ограничения

Формализовать смысл текста можно лишь при том непременном условии, что он там присутствует и выражен достаточно эксплицитно.

Общие ограничения инженерной постановки задачи:
полный анализ *предструктурированного* текста,
либо частичный анализ "информационных" текстов свободного стиля.

Отличительные черты *предструктурированного* текста (собственно "деловая проза"):

- концептуальная определенность;
- когнитивная однородность;
- тематические ограничения: ограниченная предметная область и предопределенный набор тем.

Объектом анализа могут быть

1. стилистически и лексически однородные деловые тексты, регламентированные профессиональной дисциплиной, - опирающиеся на логически и терминологически отработанную систему понятий.
2. Когнитивно однородные тексты – либо "факты", либо "законы".

(Ср.:

**Все металлы электропроводны, а вчера у нас отключили электричество.*

Исключения – общее правило и контрпример:

Зимой медведи впадают в спячку, но в нашем зоопарке медведь зимой не спит.

Проблематичны:

Метафорические контексты, смысловые пропуски – в частности, контексты, апеллирующие к энциклопедической и общекультурной компетенции читателя.

2.3. Основные подходы. Модели и методы.

1) Структурные модели.
   Уровни описания языка: фонетический / графематический, морфологический, синтаксический, семантический, прагматический. Особое место семантического уровня: не укрупнение а **переосмысление** языковых единиц.
   Семантика как междисциплинарная область.

2) Прецедентный анализ; статистический подход к языку.

"Язык описывается правилами, но состоит из исключений".

АП – авангард применения прецедентных методов (TMS)

Статистические методы как способ перехода от речи к описанию системы языка.

3) Словарная поддержка.на семантическом уровне: *онтологии*.

Nirenburg S., Raskin V. Ontological Semantics, p. 10:

Ontological semantics is a theory of meaning in natural language and an approach to natural language processing (NLP) **which uses a constructed world model, or ontology, as the central resource** for extracting and representing meaning of natural language texts, reasoning about knowledge derived from texts as well as generating natural language texts based on representations of their meaning.

#2.4. Ситуация в целом: гордиев узел проблем

Технологии полного и точного автоматического анализа делового текста пока не существует**.**

Главные проблемы:

1) Разработка и стандартизация «хорошо определенных» языков представления знаний (***ЯПЗ = KRL***) и построение систем ограниченного вывода для них.
2) Разрешение лексических и синтаксических неоднозначностей (*ambiguity resolution, disambiguation*)

   Реклама:

   *Будущее **за** окнами*

   а) пространственная интерпретация: '*будущее находится по другую сторону окон [относительно наблюдателя]* '

   б) непространственная интерпретация: '*окна имеют большие перспективы развития*' (буквальный смысл)

3) Установление референциальных отношений между единицами текста (как определить, что два разных слова в связном тексте именуют на один и тот же предмет, явление?)

Так думал ***молодой повеса***,
Летя в пыли на почтовых,
Всевышней волею Зевеса
Наследник всех своих родных.
Друзья Людмилы и Руслана!
С ***героем моего романа***
Без предисловий, сей же час
Позвольте познакомить вас:
Онегин, добрый мой приятель,
…

Ребенок был резов, но мил
…
Чтоб не измучилось ***дитя***
…

4) Теория определений и семантические примитивы (атомы смысла) в языке. (Ср. лексические функции Мельчука – Жолковского.)

5) Методы обнаружения смысловой неполноты текста и заполнения смысловых лакун.

Буквальная семантика ***vs*** косвенное выражение смысла сообщения.

<u>С. Михалков:</u>

Трусы и рубашка лежат на песке,
Никто не плывет по опасной реке.

Неполнота - одна из причин неоднозначности понимания

Посетитель в мастерской художника:
- Не можете ли Вы предложить мне что-нибудь недорогое и в масле?
- Банку сардин.

- Говорят поверхностное дыхание по Бутейко убивает вирусы гриппа.
- Может быть. Но я не представляю,как Бутейко умудряется научить эти вирусы дышать поверхностно!

Опрос таможенников бывших республик СССР – сколько времени вам нужно для покупки БМВ?
Украинский таможенник – ну, 3 месяца, не меньше.
Белорусский –месяцев 5
Российский – не менее 5-ти лет
???
Да уж больно фирма крупная.

6) Методы формализации понятийных систем. Разработка концептуальных словарей (*онтологий*), необходимых для поддержки алгоритмов семантического анализа
7) Прецедентный анализ в семантике.

Для сравнения – :

Computational semantics (IWCS-7)

January 10-12, 2007, Tilburg, The Netherlands

Endorsed by SIGSEM, the ACL Special Interest Group in Computational Semantics

TOPICS OF INTEREST

Areas of special interest for the workshop will be computational aspects of semantic theories; theoretical aspects of the design of language understanding systems and systems for multimodal communication; and semantic annotation of natural language and multimodal utterances.

TOPICS OF INTEREST:

* construction of representations of meaning in natural language
* methodologies and practices for semantic annotation
* modelling and using context in semantic interpretation
* machine learning of semantic structures
* formal and computational methods in lexical semantics

* computing meaning in multimodal interaction
* construction and use of underspecified semantic representations
* semantic concepts and ontologies
* approaches to textual entailment
* the semantics and pragmatics of dialogue acts

* the semantic web and natural language processing
* semantic aspects of language generation
* the semantics-pragmatics interface in computational perspective
* semantic relations in discourse and dialogue
* shallow and deep semantic processing and reasoning

Тема 3. Взаимодействие с синтаксическим уровнем

Формат передачи результатов синтаксического анализа должен содержать следующую информацию:

1) Исходный текст (по предложениям).
2) Выделенные лексические единицы синтаксического анализа (элементы текста) и их предварительная интерпретация.
3) Результаты синтаксического анализа (синтаксическая разметка).

Формат синтаксической разметки должен предусматривать отображение, как минимум, следующих элементов:

– числовые коды всех понятий, соответствующих слову (термину -словосочетанию);
– указание синтаксического хозяина (при локальной омонимии - всех альтернативных хозяев) и вида связи;
– выделение сегментов (части сложного предложения, обособленные обороты);
– раздельное представление всех **глобальных** вариантов синтаксического разбора;
– анафорические отсылки, распознанные парсингом;
– дополнительная грамматическая информация о слове;
– кроме того:
– термины-словосочетания;
– представление числовой информации;
– собственные имена

Типы текстовых элементов в синтаксической разметке

Тип текстового элемента	Код
Элемент, не получивший семантической интерпретации	**0**
Слово (словосочетание), выражающее указанное понятие	**1**
Число цифрами (включая запись числа в форме с порядком)	**4**
Собственное имя, идентификатор (не получившие семантической интерпретации)	**5**
Служебный элемент (союзы, знаки препинания и др.)	**7 (?)**
Дата (в стандартном формате ДД.ММ.ГГГГ) Нестандартные форматы =?	**8**

Имена синтаксических связей

Имя	Код	Описание
0_RF	255	Нет синтаксической связи
MAIN_RF	0	Главное слово (предложения или фрагмента)
NOM_RF	1	Управление именительным
GEN_RF	2	Управление родительным
DAT_RF	3	Управление дательным
ACC_RF	4	Управление винительным
INS_RF	5	Управление творительным
APP_RF	8	Приложение
ATTR_RF	9	Определительная
NIL_RF	10	Пустая связь

Имя	Код	Описание
ANAF_RF	11	Анафорическая
PGEN_RF	12	Управление родительным с предлогом
PDAT_RF	13	Управление дательным с предлогом
PACC_RF	14	Управление винительным с предлогом
PINS_RF	15	Управление творительным с предлогом
PLOC_RF	16	Управление предложным с предлогом
DMY_RF	17	Присоединяет дату

Имя	Код	Описание
SGM_RF	22	Межсегментные подчинительные связи
ANDS_RF	24	Сочинительная для сегментов
ANDN_RF	25	Сочинительная для чисел
NUM_RF	27	Подчинительная для чисел (текстовый элемент типа 4)
ID_RF	29	Подчинительная для идентификаторов (текстовый элемент типа 5)
PREP_RF	30	Отпредложная
ANDW_RF	31	Сочинительная для слов

Техника синтаксической разметки:

1) Система синтаксических связей в предложении представляется деревом зависимостей.
2) Подчинительная синтаксическая связь идентифицируется у слова – слуги ссылкой на хозяина.
3) Используются именованные синтаксические связи, номенклатура которых определена таблицей 2.
4) Сочинительные связи условно представляются как подчинительные (см. пример).
5) Сочинительные элементы (сочинительные союзы и знаки препинания) из синтаксической структуры исключаются.

Пример разметки сочинительных связей:

(1) *Красные и синие шары.*

(2) *Цветные шары и пирамиды лежат на столе.*

Вариант 1:

(1) { *(шары, синие, ATTR_RF), (синие, красные, AND_RF)* }

(2) { *(шары, цветные, ATTR_RF), (на, столе, PREP_RF),*
(шары, пирамиды, AND_RF), (лежат, шары, NOM_RF),
(лежат, на, PLOC_RF) }

Вариант 2 (представление сочинительных элементов отдельными узлами в дереве синтаксических зависисмостей):

1) { *(И, синие, ANDW _RF), (И, красные, ANDW_RF),*
(шары, И, ATTR_RF) }

4. **Синтаксическая омонимия**

1. **Виды синтаксической омонимии:**
   - Реальная – формальная
   - Локальная - глобальная
   - Омонимия адреса - содержания

2. **Омонимия разных видов связи:**
   - Омонимия подчинительных и сочинительных связей
   - Омонимия анафорических связей
   - Омонимия межсегментных связей

Явление, состоящее в том, что синтаксические связи в предложении могут быть установлены или грамматически описаны несколькими альтернативными способами.

Влечет за собой, как правило, и смысловую неоднозначность.

Реальная – формальная омонимия

Реальная:

Он из **Германии** ***туманной*** привез **учености** плоды.

Формальная: Обнаруживается, если устанавливать синтаксические связи без учета смысловых характеристик слов и / или контекста целого предложения

Лифты для высотных **зданий** ***со скоростью*** 30 м/мин.

Возьмите деревянный брусок с отверстием ***диаметром*** 30 мм.
Возьмите деревянный брусок с отверстием ***весом*** 300 г.

Мальчишек *радостный народ коньками звучно режет* ***лед.***

Еще примеры:

The plain flew over the hill. (= *над*)
The dog jumped over the fence. (= *через*)

Маркизу нельзя есть руками.

a) нельзя --(*кому?*)-- маркизу
b) есть –(*кого? что?*)--> маркизу

Локальная – глобальная омонимия

Локальная: Выбор одной из альтернативных связей для данного слова не влияет на установление связей между другими словами предложения

Глобальная: Выбор одной из альтернативных связей для данного слова влечет изменение связей между другими словами предложения

Автобус догнал трамвай

Он видел их семью своими глазами

а) Он видел
кого? семью
чью? их
чем? своими глазами

б) Он видел
кого? их
чем? своими глазами
сколькими? семью

Погибли три рабочих смены

Омонимия адреса - содержания

Омонимия адреса: Альтернативные связи по разному определяют хозяина для данного слова

Black *power struggle*

Fred saw the plane ***flying*** *over Zurich*
Fred saw the mountains ***flying*** *over Zurich*

Я ***опять*** *хочу [поехать] в Париж.*

Омонимия содержания: Альтернатива состоит в разном определении **вида связи** для данной пары «слуга – хозяин»

Выступление ***адвоката Иванова***

адвокат [чей?] – Иванова (управление)
адвокат [имеет фамилию?] – Иванов (согласование)

Омонимия разных видов связи:

- **Омонимия сочинительных связей:**
  *Вошли два человека в шляпах **и пальто.***
  *Вошли два человека в шляпах **и мальчик.***
- **Омонимия анафорических связей:**
  *Девочка уронила карандаш на пол и сломала **его.***
- **Омонимия межсегментных связей:**
  *Необходим контроль за крупными расходами граждан, которые **толкают** сегодня вверх стоимость жилья.*
- **Более сложный пример (3 варианта сочинения):**

Он постоянно видел отца, красящего забор соседа, старый дом и сарай.

1) *отец – сосед – дом – сарай;*
2) *отец – дом – сарай;*
3) *забор – дом – сарай.*

- **Омонимия семантической интерпретации синтаксической связи:**

Таблица стандартных размеров:

1) '*Таблица* ***имеет (характеристика)*** *стандартный размер'*
2) '*Таблица* ***содержит сведения о*** *стандартных размерах'*

книга сестры:

1) *' книга,* ***принадлежит*** *сестре'*
2) *' книга* ***написана*** *сестрой'*

Схема табличного представления для синтаксической разметки

Имя поля	Тип данных	Описание
Word	Текстовый	Текстовый элемент (слово или словосочетание, представляемое как узел синтаксической структуры)
NDoc	Целое число	Номер документа
FieldName	Текстовый	Имя поля документа
NSent	Целое число	Номер предложения в пределах поля
NVar	Целое число	Номер глобального синтаксического варианта
NSegm	Целое число	Номер сегмента в предложении
NWord	Целое число	Номер слова в предложении
TE	Целое число	Тип текстового элемента
Dscr	Длинное целое	Числовой код понятия
Fth	Целое число	Ссылка на синтаксического хозяина (значение NWord хозяина)
Rf	Текстовый	Имя синтаксической связи
AntS	Целое число	Номер предложения, содержащего антецедент
AntVar	Целое число	Номер глобального синтаксического варианта, содержащего антецедент
AntW	Целое число	Номер слова антецедента в указанном предложении

Пример синтаксической разметки:

Средний уровень заработной платы в Латвии вырос на 20 %, при этом уровень пенсий также увеличился.

Key	Word	ND	NS	NV	NSgm	NW	TE	Dscr	Fth	Rf	AntS	AntV	AntW
89	Средний	104	1	1	1	1	1	2001	2	ATTR RF	0	0	0
90	уровень	104	1	1	1	2	1	159	6	NOM RF	0	0	0
91	заработной платы	104	1	1	1	3	1	1523	2	GEN RF	0	0	0
92	в	104	1	1	1	4	1	901	6	PLOC RF	0	0	0
104	в	104	1	1	1	4	1	901	3	PLOC RF	0	0	0
105	в	104	1	1	1	4	1	901	2	PLOC RF	0	0	0
93	Латвии	104	1	1	1	5	1	1321	4	PREP RF	0	0	0
94	вырос	104	1	1	1	6	1	436	0	0 RF	0	0	0
95	на	104	1	1	1	7	1	908	6	PACC RF	0	0	0
96	20	104	1	1	1	8	4	0	7	PREP RF	0	0	0
97	%	104	1	1	1	9	1	204	8	NUM RF	0	0	0
98	при	104	1	1	2	10	1	916	15	PLOC RF	0	0	0
99	этом	104	1	1	2	11	1	0	10	PREP RF	1	1	0
100	уровень	104	1	1	2	12	1	159	15	NOM RF	0	0	0
101	пенсий	104	1	1	2	13	1	1524	12	GEN RF	0	0	0
102	также	104	1	1	2	14	1	0	15	NIL RF	0	0	0
103	увеличился	104	1	1	2	15	1	436	6	0 RF	0	0	0

Формат синтаксической разметки требует стандартизации ! – без чего повисает в воздухе вопрос о переносимости.

NB: Номенклатура синтаксических связей подлежит унификации!

Проект создания универсального формата разметки:

Text Encoding Initiative (**TEI**)

TEI Consortium **http://www.tei-c.org/**

Initially launched *(представлена)* in 1987,

the TEI is an international and interdisciplinary standard that helps libraries, museums, publishers, and individual scholars represent all kinds of literary and linguistic texts for online research and teaching, using an encoding scheme that is maximally expressive and minimally obsolescent.

5. Модели и методы

А. Общие подходы

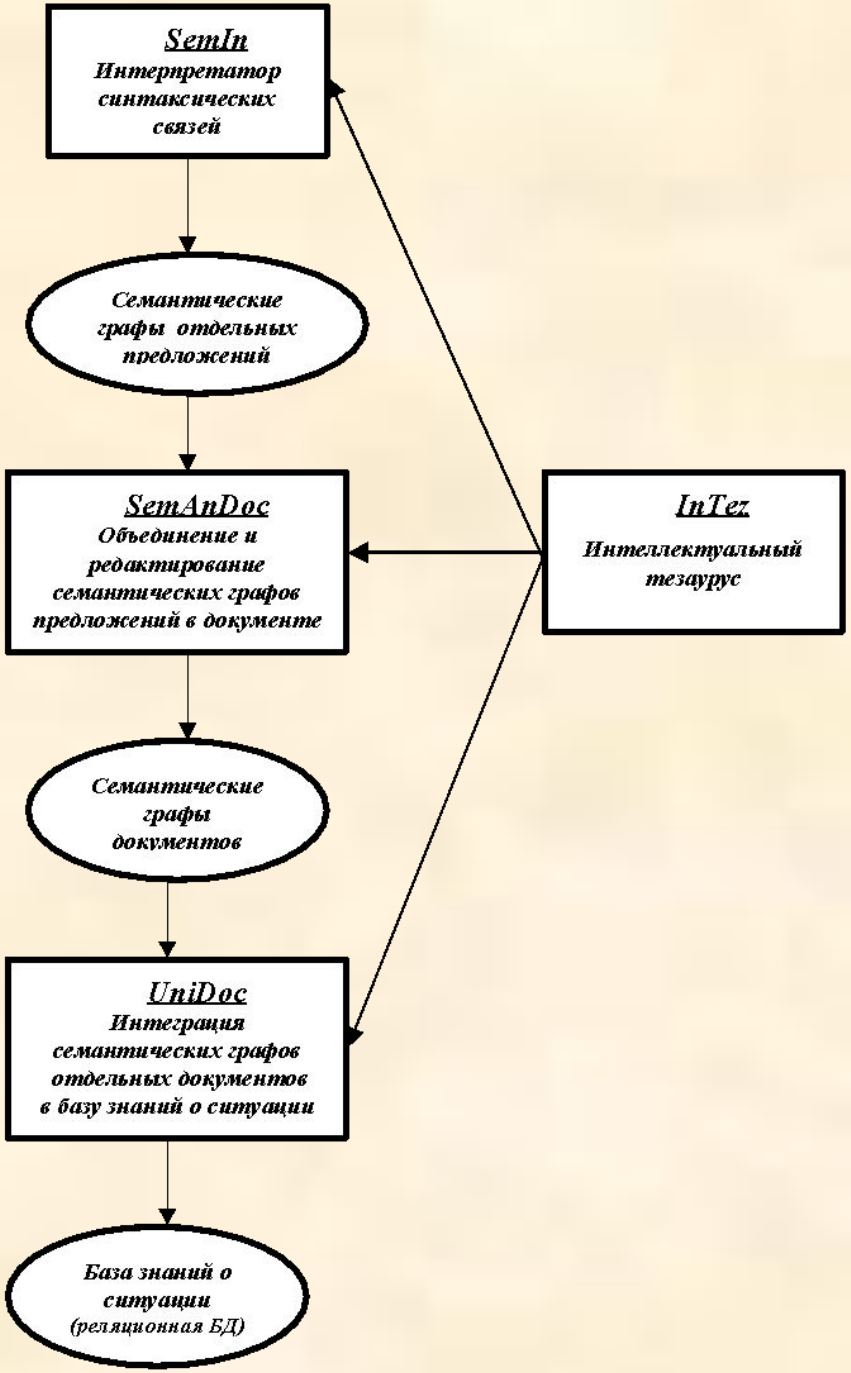
1) Универсальный целевой язык - логика предикатов.
   Другие языки (семантические сети, реляционные БД, продукционные языки) могут рассматриваться как ограниченные версии логического языка.

2) Два основных этапа анализа:
   (а) этап интерпретации грамматически выраженных (синтаксических и анафорических) связей;
   (б) этап распознавания связей не имеющих грамматического выражения.

3) В семантическом представлении лексическими единицами являются не слова, а понятия!
   Следствия:
   (а) укрупнение единиц;
   (б) размножение единиц.

4) Ключевой пункт - эффективная словарная поддержка.
   Любая система семантического анализа является тезаурусно-ориентированной.

Основная проблема в создании семантического анализатора – это проблема создания понятийного словаря, поддерживающего требуемую алгоритмами функциональность.


SemIn
Интерпретатор синтаксических связей
Семантические графы отдельных предложений
SemAnDoc
Объединение и редактирование семантических графов предложений в документе
InTez
Интеллектуальный тезаурус
Семантические графы документов
UniDoc
Интеграция семантических графов отдельных документов в базу знаний о ситуации
База знаний о ситуации
(реляционная БД)

А. Семантический интерпретатор.

Компонент, ответственный за семантическую интерпретацию грамматически выраженных связей -

как правило, в пределах предложения

(за пределами предложения – только анафора).

Предполагается, что на вход интерпретатора поступает синтаксически размеченный текст, причем в разметке сохраняются все найденные парсером варианты синтаксических связей.

В синтаксической разметке также должны быть представлены все отражаемые словарем лексические варианты (концепты) для каждого знаменательного слова.

Интерпретатор выполняет перебор и оценку предлагаемых вариантов, выбирая наиболее приемлемый (приемлемые). Таким способом в ходе интерпретации реализуется процесс разрешения лексической и синтаксической неоднозначности.

Схема переборного механизма:

[Перебор документов]
[Перебор предложений в документе]
[Перебор сегментов в предложении]

Выбор наилучшего варианта интерпретации слова или связи:

- По глобальным синтаксическим вариантам (сегментов)

-- По синтаксическим связям
(по сыновьям внутри текущего сегмента)

--- По локальным синтаксическим вариантам текущей связи
(перебор возможных хозяев для текущего сына)

---- По лексическим вариантам сына

----- По лексическим вариантам отца

Интерпретация варианта связи

к N 51

#1. Отношения, которые подлежат распознаванию

- **Ролевые:**

СООБЩАТЬ (SUB1[кто]: x_1, SUB2[кому]: x_2, OB[что/о чем]: y)

- **Кореференция**

синий шар *СИНИЙ(x) And ШАР(x)*

Предметно-ассоциативные:

дизельный автомобиль *автомобиль имеет частью дизель*

АВТОМОБИЛЬ(x) ***And*** *ДИЗЕЛЬ(y)* ***And*** *ИМЕТЬ_ЧАСТЬЮ(x, y)*

- **Функциональные:**

высокое – напряжение; весом - до - 2 - т; 200 – человек;
более - 100 – мм; 200 – мм;

- **Смысловой повтор** (смысловая избыточность):

произвел выстрел ~ выстрелил;
процесс охлаждения ~ охлаждение;
величина мощности ~ мощность;

2. Распознавание ролевых отношений

Отправным пунктом здесь является констатация того факта, что в языке имеется достаточно большой класс слов, предъявляющих определенные **требования** к контексту (как правило, требования к непосредственному синтаксическому окружению).

Такие слова принято называть **словами-предикатами**. Слово *требования* отражает точку зрения синтеза (генерации) текста.

В аспекте анализа уместнее будут слова *предсказания, ожидания.*

Требования относятся прежде всего к **смыслу** синтаксически подчиненных слов. Они регламентируют также их **возможную грамматическую форму** (падеж, предлог, возможность оформления в виде атрибутивной связи и др.)

Для приведенного выше примера:

СООБЩАТЬ (SUB1[кто]: x_1, SUB2[кому]: x_2, OB[что/о чем]: y)

Семантические требования:

SUB1[кто]: *СОЦИАЛЬНЫЙ СУБЪЕКТ (x_1)*

SUB2[кому]: *СОЦИАЛЬНЫЙ СУБЪЕКТ (x_2)*

OB[что]: *БЫТЬ_УТВЕРЖДЕНИЕМ(y)* *(?)*

Иван *сообщил* ***Петру***

но и

Правительство *сообщило всем* ***банкам…***

Иван сообщил Петру, ***что Волга впадает в Каспийское море.***

Иван сообщил Петру, ***куда впадает Волга****.*

но и

Иван сообщил Петру ***день своего приезда / о дне своего приезда***

(… что он приедет в среду)

Влияние грамматической формы предиката:

Сообщение ***Ивана*** *о* ***…*** (*GEN_RF*)

оставил сообщение ***для Петра*** (*PGEN_RF*)

но:

?сообщение ***Ивана Петру / для Петра***

Влияние лексической манифестации предиката:

Иван оповестил / уведомил / известил ***Петра***

Ср. Ожегов:

сообщить – уведомить, известить, довести до чьего-н. сведения

известить – сообщить кому-н., довести что-н. до чьего-н. сведения

Существенно, что:

1. Ожидания могут быть охарактеризованы в терминах фиксированного набора **смысловых ролей -** соответственно, можно говорить о **смысловых (семантических) валентностях,** имеющихся у слова-предиката.

2. Ожидания относятся как к смыслу, так и к грамматической форме уточняющих предикат слов.

3. **Семантические** ожидания определяются **смыслом** слова-предиката.

Совокупность таких ожиданий, описание которых хранится в концептуальном словаре, называют **семантической моделью управления** слова-предиката. **Семантическая модель управления** должна быть описана в концептуальном словаре (онтологии).

Слова-предикаты чаще всего относятся к следующим понятийным категориям.

- создание / уничтожение: *нарисовать, придумать, написать, спроектировать, построить; взорвать, разбить,* **…**
- перемещение: *приехал, падает, летит, катится, плывет; тянуть, толкать, бросить, (при)везти* **…**;
- физическое воздействие / процесс: *нагревать, резать, пилить, рвать, монтировать,* **… ;**
- Восприятие и психические процессы: *увидел, услышал, заметил, вспомнил, нашел, сосредоточил внимание на*, ;
- познавательная и коммуникативная деятельность: *узнать, догадаться, сообщить, прочитать, написать, изложить, вспомнить; сосчитать, планировать***… ;**
- биологическое поведение: *спать, болеть, питаться, схватить*, … ;

- социальное действие: *купить, приказать, арестовать, запретить, использовать, одобрять, сотрудничать,* **…** *;*
- пространственные отношения: *находиться на, внутри, снаружи, установлен на, нанесен на; вблизи, вплотную, сверху, сзади, сбоку*, **…**;
- отношения типа часть-целое: *приварен, вмонтирован, укреплен на, снабжен, содержит, состоит из,* **…**;
- социальные отношения (владения, доминирования и др.)
- отношения временной последовательности: *раньше, позже, одновременно*.
- **… …**

Толковый словарь русских глаголов: Идеографическое описание. – М., 1999.

около 25 тыс. глаголов

Отсюда – необходимость типизации описаний!

Требуют решения следующие основные вопросы:

1. Определение необходимого и достаточного набора семантических ролей (номенклатура валентностей).
2. Способы описания моделей управления у предикатных термов.
3. Способы установления соответствия между грамматической ролью имени в предложении и его семантической ролью.

Результат интерпретации:

$$R(\ldots \rho_i : x_i \ldots) \;\&\; A_i(x_i)$$

прочитал книгу □ *ПРОЧИТАЛ (… ОВ: x) & КНИГА (x)*

Описание семантической модели управления (СЕМУ):

СЕМУ ::= НОМЕР_СЕМУ <описатель валентности> |
<СЕМУ> <описатель валентности>

<описатель валентности> ::= <имя валентности> <семантическое условие заполнения> <облигат>

<имя валентности> ::= OB1 | OB2 | SB1| SB2 | INS | OB

<семантическое условие заполнения> ::=
<категориальный ограничитель> < конкретизатор>

<категориальный ограничитель> ::= <семантическая категория>

<семантическая категория> ::= НАИМЕНОВАНИЕ ПРИЗНАКА | ОБЪЕКТ | ПРОЦЕСС | ОТНОШЕНИЕ

<конкретизатор> ::= КОД_ДЕСКРИПТОРА | ПУСТО

<облигат> ::= + | -

Примеры

читать
нагреть
купить
приехать
приказать

==========

финансирование
помощь
передавать
встреча
экспорт
строительство
миграция
критиковать
использовать
обсуждать
называть

Грамматика валентностей

Семантическим моделям управления на грамматическом уровне следует сопоставлять не синтаксические модели, рассматриваемые как самостоятельные сущности, а **синтаксические условия реализации**.

Синтаксические условия реализации, вообще говоря, зависят от грамматической формы и лексической манифестации предиката:
читать – книгу (ACC_RF);
чтение – книги (GEN_RF);
прочитана – книга (NOM_RF).
сообщил (кому - DAT_RF) – *известил (кого* - ACC_RF)

Синтаксические условия реализации чаще всего определяют возможный падеж и/или предлог:

приехал – поездом (INS_RF).

приехал – на поезде (PLOC_RF / “на”).

Два пути типизации описаний

1) Типизация описаний отдельных валентностей: специфицируется семантическое условие заполнения и грамматические условия реализации.
2) Типизация СЕМУ – предикатные термы классифицируются с точки зрения возможности приписать им одну и ту же семантическую (либо семантико-синтаксическую) модель управления.

Пример - глаголы передвижения:

прибыл, отправился;

пришел, прибежал, прилетел, приплыл, приполз, **…**

Иван прилетел в Париж из Москвы самолетом Аэрофлота.

Ср.: **Иван прилетел в Париж из Москвы поездом.*

ПЕРЕМЕЩЕНИЕ (SUB1[кто]: x , OB1[откуда]: y_1 ,
OB2[куда]: y_2 ,
INS [1) способ - как; 2) средство - на чем]: z)

Рабочие гипотезы для типизации описаний:

Гипотеза 1. Для выражения основного информационного содержания научно-технического текста достаточен следующий минимальный набор имен валентностей:

OB, OB1, OB2, INS, SUB1, SUB2

Гипотеза 2 (для варианта 2). Словарь предикатных термов может быть описан конечным, и притом, обозримым списком моделей управления (несколько десятков моделей). Практически возможно разбить словарь предикатных термов на содержательные классы, соотносимые с определенным типом семантической модели управления.

Общие характеристики:

- набор валентностей;
- синтаксические условия реализации.

Класс '*физическое воздействие на материал'* (SUB1, OB, INS):
нагревать, строгать, пилить, сжимать
vs *коррозия*

Возможная синтаксическая роль актанта определяется грамматикой ролевых связей, устанавливающей соответствие вида

(Rf , GFP, TSEMU) --> VAL_,

где

Rf - имя синтаксической связи;

GFP - грамматическая форма предиката,;

TSEMU - семантико-синтаксический тип предиката (словарная характеристика – предполагается типизация актантных структур!);

*VAL_* - имя возможной валентности, либо отсылка к ролевой функции предлога.

Для предложных связей проверяется словарно определяемая способность предлога служить указателем роли для падежа, указываемого синтаксической связью Rf.

Дополнительно проверяется соответствие актанта семантическому условию заполнения валентности предиката (проверка на объемную совместимость).

Грамматика ролевых связей – языково-зависимый компонент.

Может быть реализована в форме внешней таблицы –

что должно обеспечить настройку на язык входного текста без корректировки кода.

(Возможный вариант реализации - компилируемая таблица.)

RF GFP TSEMU VAL ПРИМЕРЫ

==========

NOM_RF VA 14 SUB1 *Россия в 2001г.* ***продала*** *развивающимся странам оружия на сумму 5,7 млрд;*

NOM_RF VP 14 OB *товары*, ***поставляемые*** *из КНР;*
HO:

GEN_RF NV2 OB ***нагревание*** *воды;*

GEN_RF NV3 OB1 ***сварка*** *меди (с…)*

GEN_RF NV8 OB ***коррозия*** *металла*

GEN_RF NV14 OB ***экспорт*** *(импорт, покупка, продажа, поставка) реактивного топлива*

RF GFP TSEMU VAL ПРИМЕРЫ

==========

DAT_RF VA 14 SUB2 *Россия в 2001г.* ***продала*** *развивающимся странам оружия на сумму 5,7 млрд.*

ACC_RF VA 14 OB *Казахстан* ***закупит*** *новые истребители*

ACC_RF VA 0 SUB2 ***встретил*** *друга;*

ACC_RF VA 0 OB1 ***нагрел*** *воду*

INS_RF VA 5 SUB2 ***руководит*** *отделом*

INS_RF VA 5 OB ***управляет*** *самолетом / плавкой*

INS_RF NV 15 SUB2 ***руководство*** *отделом*

INS_RF VP 4 SUB1 ***перевозится*** *фирмой*

INS_RF VP 4 INS ***перевозится*** *самолетами*

RF GFP TSEMU VAL ПРИМЕРЫ

==========

PACC_RF VP 14 PREP *боевые самолеты марки "СУ"*
поставлялись *<u>в Индию</u> (SUB2)*

PACC_RF VA 4 PREP ***прилетет****ь <u>на Сахалин</u> (OB2);*

PLOC_RF VA 6 PREP *изготовить на станке (INS)*

PLOC_RF VA 6 PREP ***приехать*** *<u>на поезде</u> (INS)*

3. Распознавание отношения контактной кореференции

Различительный тест - возможность синонимических трансформаций словосочетания – в том числе с изменением направления синтаксической связи.

синий шар = (-)

шаровая молния = **молниевый шар / молния в форме шара*

жидкий диэлектрик = *диэлектрическая жидкость*

магниевый порошок / порошок магния = *порошковый магний*

аморфный кремний = (-)

кристаллический кремний = *кремниевый кристалл / кристалл кремния*

медные листы = *листовая медь*

металлический куб = ?*кубический металл / металл в форме куба*

Общая логическая схема интерпретации:

$$P_F(x) \ \& \ P_S(x)$$

или

$$P_F(x, v_F) \ \& \ P_S(x, v_S)$$

Для установления контактной кореференции необходимы и достаточны условия:

1) Хозяин и слуга принадлежат семантической категории *Объект*.
2) Понятия, соответствующие термам хозяина и слуги, находятся в отношении объемной совместимости.
3) В случае предложной связи - способность предлога выражать отношение кореференции

   (НО: *посуда <u>из</u> стекла* **vs** *посуда <u>из</u> Чехии*).

Данная гипотеза может быть распространена на все виды десемантизированной подчинительной связи, такой как связи типа ***A + N*** (прилагательное + существительное)
и ***N + N***$_{GEN}$ (управление беспредложным родительным) в русском языке;
связь типа ***N + N*** в английском языке *(magnesium powder),*

и т.д.

4. Распознавание функциональных отношений

- признак – значение признака:
  высокое – напряжение;
  весом - 2 [m]
- число – единица измерения; число – имя объекта:
  200 – мм;
  200 - человек
- число – модификатор значения:
  более - 100 - мм
- терм - отрицание
  200 - мм
- логический оператор – соединяемые термины:
  *синий **И** красный шары*

4.1. Анализ количественных групп.

Что такое количественные группы?

Стандартный пример:

Жесткие диски емкостью до 100 ГБ.

Основные элементы:

1) имя объекта: *жесткие диски;*
2) наименование признака: *емкость;*
3) количественное значение: *100;*
4) единица измерения: *ГБ*
5) модификатор значения: *до.*

Некоторые из элементов могут отсутствовать:

Жесткие диски до 100 ГБ.

Виды количественных значений и их представление:

А. числовые

1) точечные:

мощностью 100 вт $МОЩНОСТЬ\_вт(x, v) \ \& \ v = 100$

2) интервальные:

- зона, ограниченная снизу: *мощностью свыше 100 вт*;
- зона, ограниченная сверху: *мощностью до 100 вт*;
- собственно диапазон:

мощностью от 100 до 1000 вт

$$МОЩНОСТЬ\_вт(x, v) \ \& \ v >= 100 \ \& \ v <= 1000$$

3) представляющие числовую оценку динамики изменения:

- «на сколько» - абсолютная оценка:

мощность увеличена на 100 вт;

$$МОЩНОСТЬ\_вт\ (x, v)\ \&\ Увеличение\_на\ (v, 100)$$

- «во сколько»: *мощность выросла в 1,5 раза*;

$$МОЩНОСТЬ\_вт\ (x, v)\ \&\ Увеличение\_в\ (v, 1{,}5)$$

- «на сколько» - относительная оценка:

мощность упала на 20 %.

$$МОЩНОСТЬ\_вт\ (x, v)\ \&\ Уменьшение\_на\_\%\ (v, 20)$$

Б. нечисловые

2) нормативно-оценочные: *большой мощности;*

$$МОЩНОСТЬ\_вт\ (x, v)\ \&\ БВ\ (v)$$

2) представляющие динамику изменения оценочно-вербально:

мощность растет

$$МОЩНОСТЬ\_вт\ (x, v)\ \&\ Увеличение\ (v)$$

Задачи, решаемые анализатором:

1) Разграничение ***величин*** и ***количеств***:

20 человек **vs** *20 м*

2) Интерпретация именованного числа как значения признака; пересчет значения к стандартной единице измерения

10 квт *10 000 вт (мощность)*

3) Присваивание признаку значения; уточнение наименования признака:

толщиной 100 мкм (признак *линейный размер* уточняется как *толщина*)

4) Преобразование вербальных и вербально-цифровых значений в числовой формат; восстановление сокращенных обозначений элементов числа

тысяча сто *1100*

10 млн. *10 000 000*

5. Смысловой повтор

Отношения **смыслового повтора** обнаруживаются в словосочетаниях, обладающих смысловой избыточностью:

произвел выстрел ~ выстрелил;
осуществил расчет ~ рассчитал;
процесс охлаждения ~ охлаждение;
отношение предшествования ~ предшествование;
величина мощности ~ мощность;
красного цвета ~ красный.

#6. Предметно-ассоциативные отношения

Связь между синтаксическим хозяином и слугой допускает конкретную содержательную интерпретацию; словосочетание может быть трансформировано в синонимичную трехчленную конструкцию, в которой связь получает явное лексическое выражение термином, представляющим некоторое отношение предметной области:

дизельный автомобиль автомобиль имеет частью дизель;

учебный автомобиль автомобиль используется для обучения;

радиационная проводимость проводимость имеет причиной

радиацию;

продуктовый магазин магазин, торгующий продуктами;

цистерна с нефтью цистерна, содержащая нефть.

При такой интерпретации различимы следующие смысловые составляющие:

(1) дескрипция B(y), соответствующая синтаксическому хозяину;

(2) дескрипция A(x), соответствующая синтаксическому слуге;

(3) подразумеваемое (не имеющее лексического выражения в тексте) отношение R, устанавливаемое между сущностями, указанными референциальными индексами x и y.

Соответственно, получаем следующую логическую схему интерпретации:

$$A(x) \,\&\, B(y) \,\&\, R(x, y)$$

Выбор «предметного» отношения при такой интерпретации может быть мотивирован по-разному.

Для связей, маркируемых предлогом, одна из возможных мотивировок - указание отношения самим предлогом.

рукопись на столе → рукопись находится_на столе;

рукопись в столе → рукопись находится_внутри стола;

рукопись под столом → рукопись находится_под столом;

Здесь именно предлог (для русского - взятый вместе с падежом управляемого слова) определяет выбор подразумеваемого отношения.

Информация о потенциальных возможностях предлога выражать в определенных контекстах то или иное предметное отношение также должна присутствовать в словаре.

Для связей, НЕ маркируемых предлогом - может определяться тезаурусным отношением между концептами сына и отца.

Для установления **специфицируемых предметно-ассоциативных отношений** необходимы и достаточны условия:

1) Понятия, соответствующие термам хозяина и слуги, находятся в отношении объемной несовместимости, либо (в случае совместимости) эти термы синтаксически связаны через предлог, не способный выражать отношение кореференции.
2) С парой термов *хозяин – слуга* словарно ассоциировано некоторое предметное отношение

(*<автомобиль, кузов> --> иметь частью*)

книга издательства, книга сестры, книга анекдотов,…

и/или (если связь предложная) предметное отношение ассоциировано с предлогом и падежом.

Для установления **не специфицируемых предметно-ассоциативных отношений** необходимым и достаточным является истинность первого и ложность второго условия.

Таким образом, при описании предлогов в семантическом словаре следует предусмотреть ответы на следующие вопросы:

(1) какие роли при предикатном терме может маркировать данный предлог;

(2) может ли он маркировать связь кореференции;

(3) какие лексические (предметные) отношения он может выражать;

(4) на какие ограничения или функции числовых величин он может указывать.

Б. Основные постулаты интерпретации синтаксических связей.

1) Тип устанавливаемого семантического отношения определяется семантическими характеристиками хозяина и слуги.

   Соответственно, работа интерпретатора должна управляться категориальной принадлежностью членов интерпретируемой связи.

   Грамматическое оформление синтаксической связи – в одних случаях будет учитываться при определении *конкретного содержания* семантического отношения (например, выбор конкретной валентности или терма для предметно-ассоциативного отношения), в других (и достаточно многочисленных!) случаях вовсе не играет роли.

2) Интерпретация синтаксической связи является контекстно-свободной.
3) Предлоги рассматриваются не как самостоятельный объект интерпретации, а как дополнительная (семантико-грамматическая) характеристика связи между синтаксическим хозяином предлога и управляемым предлогом знаменательным словом.
4) Лексические и локальные синтаксические неоднозначности (наличие у слова альтернативных хозяев) обрабатываются в одном переборном механизме. При этом используется система эмпирически устанавливаемых предпочтений.

NB: Никаких специальных алгоритмов разрешения неоднозначностей в такой модели не используется!

(Глобальные варианты синтаксического разбора предложения рассматриваются в переборном механизме следующего уровня. В этом случае сравниваются *суммарные веса интерпретации всех связей* предложения.)

Порядок просмотра связей в синтаксическом графе именной группы процедурой семантической интерпретации, вообще говоря, имеет значение. Правильный результат можно получить, если вести просмотр снизу вверх (от подчиненных к подчиняющим) с использованием при проверке совместимости накопленной информации об объекте-референте.

Так, например, в конструкции ***ротор с переменным диаметром вала*** просмотр сверху вниз даст неправильный результат: объект *ротор с переменным диаметром* связан с объектом *вал* (ср. сходную конструкцию *вал с переменным диаметром,* где такой анализ будет правильным).

Связи согласования (определительные связи) при данном хозяине должны интерпретироваться прежде, чем связь управления.

Порядок предпочтений при выборе "наилучшей" интерпретации:

- функциональные связи и связи, устанавливающие факт смысловой избыточности;
- ролевые – при наличии семантически согласованного актанта;
- связи кореференции;
- ролевые связи, определяемые как факультативные или не подтвержденные семантическим согласованием;
- предметно-ассоциативные связи специфицируемые;
- предметно-ассоциативные связи не специфицируемые;
- отсутствие тезаурусных связей.

В случае обнаружения синтаксической омонимии ***сочинительных*** связей предпочтения определяются степенью согласованности семантических характеристик участников синтаксической связи.

Примеры:

1) *Экспериментатор воздействовал на* ***спины*** *элементарных частиц.*

Онтология:

а) ‘Элементарная частица’ характеризуется признаком ‘*спин*’

Логическая интерпретация:

СПИН (x, v) & ЭЛЕМЕНТАРНАЯ_ЧАСТИЦА (x) &

ВОЗДЕЙСТВИЕ (Sub1:y, Ob: v) & ЭКСПЕРИМЕНТАТОР (y)

б) ‘Хордовые животные’ [*они и только они*] имеют частью ‘*спину’*

б') Концепты ‘Хордовое животное’ и *‘Элементарная частица’*

- объемно несовместимы

Общий подход (для лексической неоднозначности)– учет трех типов факторов [Agirre E., Stevenson M., WSD, p.p. 224 - 228]:

1) свойства самого слова;
2) свойства локального контекста;
3) свойства глобального контекста.

Контрпримеры:

(1) *Эти типы стали есть в прокатном цехе.*

Возможные средства разрешения (NB: алгоритм должен обнаружить проблему!):

- подсчет суммарной оценки качества интерпретации для предложения;
- *типы* – разг. стиль;
- общий контекст (производственный?; о стали уже шла речь?); семантическая "когерентность" предложения предшествующему тексту (вопрос о мере);
- статистика сочетаемости - *есть* в значении *принимать пищу* и *цех* – редко вместе?

(2) *The **box** was in the **pen**.*

Bar-Hillel (1964)

Невозможность использования основных значений:

**Коробка была/находилась в пере/ручке.*

Необходимость обращения к предшествующему контексту – какие из предметов, указанных в толкованиях, ранее упоминались?

Словарь Контекст 6.0:

pen *n*

1. перо
   (писчее)
2. ручка
   (для письма - с пером, авторучка, шариковая и т.п.)
3. рейсфедер
   (чертежный)
4. литературный стиль
5. писатель

---

6. небольшой загон
   (для скота, птицы)
7. небольшая огороженная площадка
   (и т. п.)
8. плантация, ферма
   (на Ямайке)
9. помещение для арестованных
   (при полицейском участке)

   ---

10. самка лебедя, лебедка

box n

1. коробка, ящик, сундук.
2. рождественский подарок (обычно в ящике)
3. ящик под сиденьем кучера
4. козлы
5. театр. ложа
6. стойло
7. маленькое отделение с перегородкой (в харчевне)
8. домик (особ. охотничий)
9. рудничная угольная вагонетка
10. тех. букса
11. вкладыш
12. втулка
13. бокс
14. удар !!!
15. бот. самшит вечнозеленый

---

box v **…**

Компьютерный спецсловарь в коплекте Контекст 6.0:

box n

1. стойка, шкаф
2. блок
3. прямоугольник, рамка, окно, управляющее окно

Изображение прямоугольника на блок-схеме, графике или экране дисплея.

box

блок, модуль, стойка

(3) *The **astronomer** married the **star**.*

Charniak (1983)

Невозможность использования критерия предметной области.

Необходимость обращения к модели управления концепта '*marry*'^

Словарь Контекст 6.0:

star n

1. звезда, светило
2. звезда, ведущий актер или актриса; выдающаяся личность
4. полигр. звездочка
5. звездочка (белая отметина на лбу животного)
6. нечто , напоминающее звезду
7. судьба, рок
8. ведущий

---

star adj

1. звездный
2. выдающийся
3. великолепный

---

star v …

2. Распознавание связей, не имеющих грамматического выражения.

Основная проблема - кореференция имен объектов.

Примеры:

1) *Так думал молодой **повеса… Наследник** всех своих родных. . . С **героем** моего романа. . . **Онегин**, добрый мой приятель. . . Судьба **Евгения** хранила... **Ребенок** был резов, но мил.*

2) *Вот бегает дворовый **мальчик**, / В салазки жучку посадив,*
*Себя в коня преобразив. **Шалун** уж заморозил пальчик**…***

3) *Кампоманес не склонен терять время на попытки вернуть Фишера на шахматную арену**…** Прошло уже двенадцать лет, как победитель матча в Рейкъявике оставил шахматы.**

3) *Недавнее землетрясение самым пагубным образом отразилось на **Венеции**. . . **Уникальный исторический центр** может выжить лишь при условии, что итальянское правительство примет самые срочные меры по устранению угрозы затопления **города** водами Адриатики.*

4) *Эффективность красной люминесценции **фосфида галлия**. Проведены исследования оптических свойств **кристаллов**.*

5) *Итальянское правительство заключило с правительством России соглашение о сотрудничестве в области энергетики.*

6) *Слава богу! **Грозненский «Терек»** наконец-то проиграл и выбыл из кубка УЕФА. Впервые за сорок лет болельщицкого стажа я радуюсь проигрышу **отечественного клуба** иностранной команде. Надоело наблюдать, как наши телеканалы делают из совершенно рядового события – участия **заштатного футбольного клуба** в первой стадии международного турнира – политическое событие едва ли не всероссийского масштаба.*

Примеры кореферентных связей (по *Nirenburg & Raskin*)

Direct reference by name:
Last week Bill Clinton went on an official visit to Turkey,
Greece and Kosovo.

Pronominalization and other deictic phenomena:
The goal of his visit to these countries was to strengthen their ties
with the United States.

Indefinite and definite descriptions:
This was the President's first trip to the Eastern Mediterranean.

Ellipsis:
He traveled [*to Turkey, Greece and Kosovo* - elided] *by Air Force One.*

Non-literal language (that is, metaphors, metonymies and other tropes):
The White House chief (metonymy) *hopes that the visit*
will stem the tide (metaphor) *of anti-American protests in Greece.*

Примеры построения связного текста:

(1-1) ***Авианосец*** *"Йорктаун" получил большие повреждения и был затоплен.*

(1-2а) **…** *Крейсера повреждений не получили.*

(1-2б) **?** **…***Корабли повреждений не получили.*

/ + другие, остальные, **…**/

(2-1) Завод "Электросила" производит мощные электрические машины.

(2-2) [*Аналогичное*] *предприятие находится в Харькове.*

Общие соображения:

1) **Тотальность** задачи анализа референции для любого текста.

   В лингвистических работах сравнительно недавнего прошлого кореференция (анафора) связывалась лишь с некоторыми достаточно специфичными средствами выражения смысла (такими как местоимения и лексический повтор). Сейчас осознан (в вычислительном аспекте) ее универсальный характер.

   Построение семантического представления текста предполагает в качестве основной процедуры приписывание **каждому** знаменательному слову с предметным значением референциального индекса.

   Это значит, что процедура анализа для каждого такого слова должна либо произвести выбор одного из уже имеющихся в семантическом представлении референциальных индексов, либо открыть новый индекс.

2) Анализ кореференции актуален как при рассмотрении дистантных (в частности, межфразовых связей), так и при рассмотрении связей **в пределах простого предложения**, и прежде всего - связей непосредственного синтаксического подчинения.

3) Следует различать собственно лингвистические описания и возможность реализации этих описаний в моделях анализа.

   На описательном уровне собран большой и разнообразный материал; дело за тем, чтобы привести точки зрения разных авторов в единую систему. На уровне моделей анализа пока либо рассматривается весьма ограниченный круг явлений, либо высказываются содержательные соображения, способы и средства алгоритмизации которых до конца не ясны.

4) Весьма ограниченная применимость прецедентных методов.

Для анализа отношений кореференции в пределах простого предложения наиболее значим учет актантной структуры предложения.

При большинстве предикатов сопредикатные имена должны обозначать **разные** объекты, т.е. не могут быть кореферентны.

Рыбак рыбака *видит издалека.*

Ворон ворону *глаз не выклюет.*

Исключение — предикаты кореференции:

Экран *изготовлен из* ***меди****.*

В качестве ***внешней памяти*** *используются* ***видеодиски****.*

Гипотеза индикации - концептуально простая модель, опирающаяся на словарный механизм вычисления объемной совместимости имен.

Исходное предположение:

При построении (понимании) текста существенно используется информация о совместимости (несовместимости) предметных имен. Эта информация полагается априорной относительно процедуры анализа (синтеза) данного текста ("тезаурус", которым обладает человек или система понимания, воспринимающая либо порождающая текст)

Референциальное отождествление имен объектов в связном тексте определяется тремя факторами:

- порядком следования имен в тексте;
- совместимостью / несовместимостью имен;
- наличием индикаторов референции.

Для несовместимых имен нулевой индикатор маркирует референциальное различие,

для совместимых - референциальное тождество.

Содержание гипотезы индикации весьма компактно может быть представлено в табличной форме. Таблица отражает точку зрения анализа текста (на входе — сведения о маркированности второго имени и о совместимости имен, на выходе — решение о необходимости референциального отождествления имен). Символы $= (\neq)$ означают, что при данной комбинации условий имена получают один и тот же (разные) референциальные индекс; φ - признак совместимости (1 — ДА, 0 - НЕТ).

Индикатор референции	I^+	I^0	I^-
Имена совместимы	$=$	$=$	$\neq$
Имена несовместимы	$\neq$	$\neq$	$\neq$

Прецедентный анализ.

Анализ "по образцу" (example-based, case-based,…), основанный на использовании корпуса предварительно размеченных текстов.

Пока - большие надежды и много проблем.

Формат семантической разметки текстов?

Поддержка функциональностью семантического словаря (генерализация образцов) более чем актуальна.

> *Средняя зарплата оказалась больше на 1000 руб.*
> *Полетный вес будет уменьшен на 0,5 т.*

Проблема накопления корпуса образцов – как побочный результат работы анализатора с постредактированием.

Словарная поддержка процедур семантического анализа

"Семантический анализ – это словарь!"
(Процедуры семантического анализа во всех без исключения случаях опираются на функциональность понятийного словаря.)

Проект ***Shalmaneser*** (a SHALlow seMANtic parSER):

"One of the most urgent problems (*острых проблем*) in language technology is the lexical semantics bottleneck, the unavailability of domain-independent lexica with rich semantic information on lexical items. Such lexica could greatly improve the quality of current applications. At the same time, providing large-scale lexical semantic information is an enormous challenge, due to the size of the vocabulary and the inherent vagueness of lexical meaning."

Ключевые моменты:

1. Должна быть четко различена лингвистическая и концептуальная лексикография. Словарь для поддержки семантического анализа должен описывать свойства и отношения **понятий, а не слов**. Любые словари, ограничивающие себя рассмотрением отдельных слов, окажутся мало полезными для такого применения.
   Концептуальная лексикография конституируется дисциплинарно как *вычислительная онтология*.

2. Точнее, нужны **два** словаря: кроме собственно концептуального словаря нужен **словарь перевода**, определяющий соответствие *слова* <--> *понятия*. Часто словарь перевода совмещается со словарем основ.

3. Концептуальный словарь должен представлять собой нечто большее, чем просто таксономию. Для моделей анализа ключевыми являются следующие функции:

- детальная семантическая категоризация лексики;
- вычисление полного набора объемных отношений (*включение – совместимость – несовместимость*);
- определение возможных для заданной пары понятий предметно-ассоциативных отношений:
- описание семантических моделей управления предикатов;
- для отдельных семантических классов - задание узко специальных связей (понятие '*красный*' дает ответ на вопрос о *цвете* вещи, а понятие '*горячий*' – не дает; *мощность* может измеряться *ваттами*, но не *тоннами* и т. д.)
  NB: Описание семантики предлогов!

Словарь или словари?

Можно ли создать концептуальный словарь как единый унифицированный вычислительный ресурс (*sharable and reusable - T. R. Gruber*)?

Зачетные задания:

1. см