Проектирование человеко-машинных интерфейсов

Лекция 9. Моделирования понимания естественного языка

НГТУ, кафедра ВТ
А.В.Гаврилов

«Понимающий» компьютер должен осуществлять обработку всех уровней языка

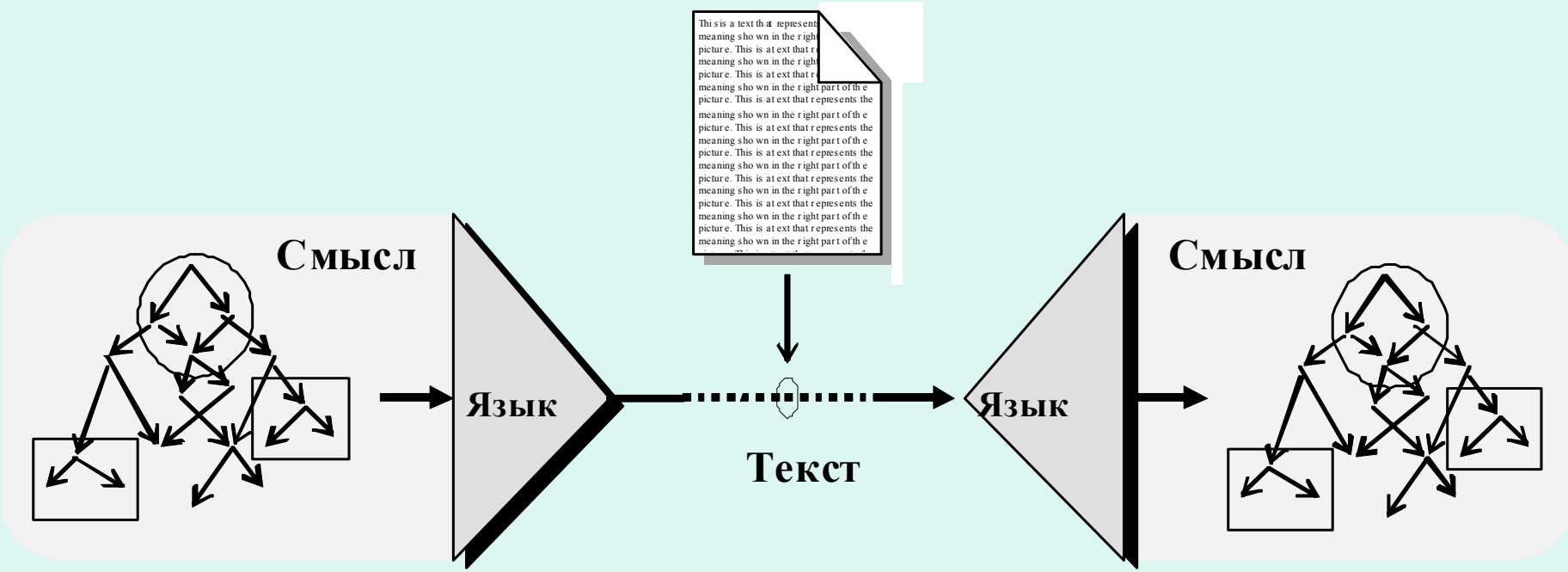
- Фонетика (при голосовом общении)
- Морфология
- Синтаксис
- Семантика
- Прагматика
- Дискурс

Решается только одна проблема:

НЕОДНОЗНАЧНОСТЬ

ОСОБЕННОСТИ ЕЯ: ПРЕОБРАЗОВАТЕЛЬ СМЫСЛ-ТЕКСТ

- Объект – *текст*
- Линейность текста
- Составлен из различных единиц
- Единицы принадлежат к разным уровням



ОСОБЕННОСТИ ЕЯ: УРОВНИ и ПОДУРОВНИ

- Синтаксический (предложения ЕЯ)
  - подуровень словосочетаний (*увидел лес*, *красивый закат*)
  - надуровень сверхфразовых единств (сложных синт. целых ≈ абзацев), объединяющихся по смыслу и лексико-грамматически (повторы слов, анафорические ссылки)
- Морфологический (слова ЕЯ, словоформы)
  - Подуровень морфем; м*орфема* – минимальная значимая часть слова (корень, приставка, суффикс…)
- Фонологический (звуки / символы)

? Уровни/ Срезы ?

- Семантический - набор элементарных единиц – *сем*
- Лексический: *лексема* – совокупность *словоформ* слова (*конь*, *коня*, *коню*, *коне*)
- *Дискурсивный* (связный текст) – схематические структуры текстов (патентные формулы, деловые письма и т.п.)

ЕЯ и ИСКУССТВЕННЫЕ ЯЗЫКИ

Искуств. языки, например: языки программирования
Близки по функциям, но

Принципиальные отличия:

- Открытость и изменчивость ЕЯ (на всех уровнях) ⇒ невозможность единожды разработать лингв. процессор
- Нестандартная сочетаемость (*синтактика*) единиц ЕЯ на всех уровнях, например, *лексическая* сочетаемость:
  крепкий чай, но не *тяжелый чай* (*heavy tea*)
- Большая системность (число уровней) и степень ассиметрии связи единиц и выражаемых ими смыслов
  - Полисемия (многозначность)
  - Синонимия (совпадение смыслов)
  - Омонимия (совпадение форм)

МОДЕЛИРОВАНИЕ В КЛ

Модель языка – описание свойств обрабатываемого текста.

Особенности моделей КЛ:

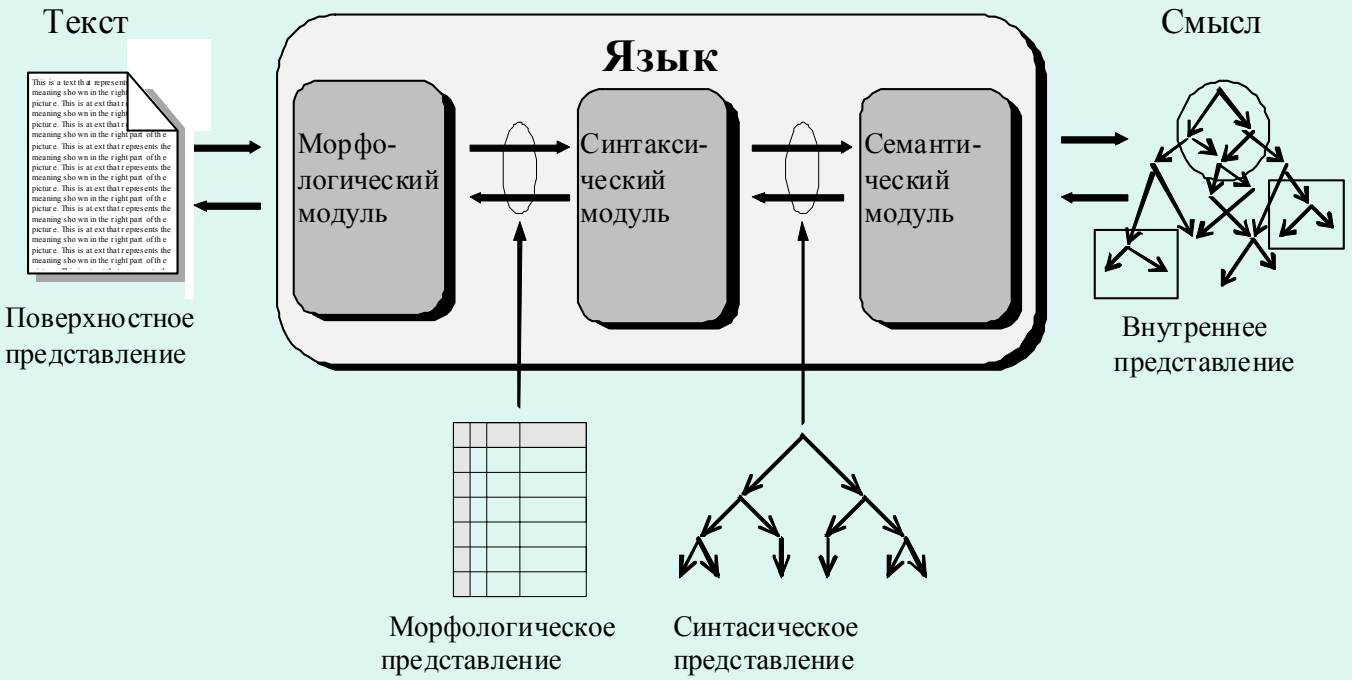
- Формальность и алгоритмизируемость;
- Функциональность: цель – воспроизведение функций языка как «черного ящика», а не моделирование языковой деятельности человека;
- Общность модели, т.е. покрытие ею довольно большого множества текстов;
- Экспериментальная обоснованность, предполагающая тестирование модели
- Опора на те или иные словари как обязательную составляющую модели.

МОДУЛЬНОСТЬ ЛИНГВ. ПРОЦЕСОРОВ

Сложность ЕЯ $\Rightarrow$

лингвистический процессор – многоэтапный преобразователь

- Анализ текста: первичный модуль – графематический анализ
- Синтез текста: другое направление обработки



ВИДЫ И ОСОБЕННОСТИ МОДЕЛЕЙ

В зависимости от учета уровней ЕЯ:

- Структурные (несколько уровней)
- Редуцированные - Статистическая модель : статистика символов/букв, их биграмм и триграмм (уровень символов)
  или слов, их биграмм и триграмм
- Структурно-статистические

На разных уровнях ЕЯ:

- Модели морфологии (анализ: лемма или основа с морфологическими характеристиками исходной словоформы)
- Модели синтаксиса, анализ: синтаксическое дерево:
  - деревья непосредственно составляющих (валентности слов, например: *передать - кто? кому? что?* – subcategorization frame)
  - деревья зависимостей (валентности – модели управления слов)
- Модели семантики представление смысла (свойства, отношения, состояния, действия) – на основе моделей ИИ: формулы исчисления предикатов или семантические сети

МОДЕЛЬ «Смысл⇔Текст»

И. А. Мельчук, Ю. Д. Апресян (с 70-х годов)

Смысл – инвариант синонимичных преобразований текста.

- ориентация на синтез текстов
- многоуровневость модели, разделение основных уровней на поверхностный и глубинный уровень, например: *глубинный* (семантизированный) и *поверхностный* («чистый») синтаксис.
- Сохранение всей информации при переходе с уровня на уровень;
- *Лексические функции* для описания нестандартной синтактики, на их основе сформулированы правила синтаксического перифразирования;
- Упор на словарь, а не на грамматику; в словаре – информация для разных уровней языка (синтаксис: модели управления слов, описывающие их синтаксические и семантические валентности);
- Семантическое представление текста: семантический граф + коммуникативная организация смысла

ЛИНГВИСТИЧЕСКИЕ РЕСУРСЫ

Лингвистические процессоры базируются на определенном представлении лингвистической информации:

- Компьютерные словари
- Грамматики ЕЯ
- Базы словосочетаний
- Тезаурусы и онтологии
- Коллекции и корпуса текстов

ЛИНГВИСТИЧЕСКИЕ РЕСУРСЫ: СЛОВАРИ и ГРАММАТИКИ

Словари для ЛП обычно разрабатываются специально .

Различаются:

- Охватом лексики: общая/специальная
- Представленной информацией (в словарной статье):
  - морфологические словари
  - словари моделей управления
- Видом:
  - словари синонимов:
  - словари паронимов: *чужой* и *чуждый*, *правка* и *справка*
  - словари терминов некоторой предметной области

Грамматики – набор правил, описывающих структуру предложений:

Пример:

SUBJECT|gender 1 ^, number 1 ^, case 1 ^|<1:;SBJ1;gender 1 +,number 1 +,case 1 +>|<1:;SPRE;gender 1 +,number 1 +, case 1 +>|<1:;SPOST;gender 1 +,number 1 +, case 1 +>

ЛИНГВИСТИЧЕСКИЕ РЕСУРСЫ: БАЗЫ СЛОВОСОЧЕТАНИЙ

Сравнительно новый тип лексического ресурса,

Отражает стандартную и нестандартную сочетаемость слов ЕЯ

Обширная база словосочетаний РЯ – система **КроссЛексика**

- Примерно миллион словосочетаний общей лексики
- Словосочетания многих синтаксических типов:
  - определяемое слово → определитель (*полевая форма*, *вполне удачный*)
  - существительное → его дополнение (*рост возмущения*)
  - глагол → его дополнение (*заметить разницу*, *решить продать*)
  - прилагательное → его дополнение (*дошедший до ручки*)
  - сочиненная пара (*наземный и воздушный*, *орел и решка*)
- Семантические связи слов: синонимы, антонимы, гиперонимы, холонимы
- Пометы стиля слов (устарелый, разговорный, бранный, и т.п).

ЛИНГВИСТИЧЕСКИЕ РЕСУРСЫ: ТЕЗАУРУСЫ И ОНТОЛОГИИ

- Тезаурус – семантический словарь
  - **РуТез** – информационно-поисковый тезаурус, 52 тыс. понятий из общественно-политической области; связи: синонимия, род-вид (выше-ниже), ассоциация, онтологическая зависимость,
  - **КроссЛексика** (поскольку представлены смысловые отношения)
- Онтология – формальное описание определенного набора понятий, сущностей
  - **WordNet** – лингвистическая онтология на базе английских слов
    - Дж. Миллер, 1984 г., модель человеческой памяти
    - слова разбиты по частям речи
    - для слов каждой части речи выделены *синсеты* – наборы синонимов
    - версия 3.0 – 155 тыс. лексем, 117 тыс *синсетов* (понятий)
  - **EuroNet** – аналогичные лексические ресурсы для других европейских языков

ЛИНГВИСТИЧЕСКИЕ РЕСУРСЫ: КОРПУСА ТЕКСТОВ

Трудоемкость создания лингвистических процессоров и лексических ресурсов $\Rightarrow$ автоматизация их построения

- *Коллекция текстов:* представительный набор текстов, собранных по определенному принципу
- *Корпус текстов:* коллекция текстов с лингвистической разметкой: морфологической, лексической, синтаксической, дискурсивной

  - использование в лингвистических исследованиях

  - применение для машинного обучения моделей

  - для РЯ – Национальный корпус русского языка
- *Интернет-корпус:* тексты сети Интернет как корпус современной речи

Два подхода к моделированию понимания смысла ЕЯ

- *Синтаксически-ориентированный подход* основан на детальном синтаксическом разборе предложения. Средствами синтаксического анализа вычленяются связанные понятия , которые объединяются в так называемые атомы смысла (АС). Создание АС идет только на основе данного предложения , определение связанный понятий идет только на основе синтаксических правил.
- *Семантически-ориентированный подход* на основе распознавания семантики. Разбор предложения идет путем вычленения связанных понятий с помощью базы знаний. База знаний хранит АС и определенным образом представленные связи между этими АС. На втором плане стоит синтаксический анализ с помощью которого вычленяются дополнительные АС, те которые не были сгенерированы с помощью БЗ.

Технологии анализа ЕЯ. Синтаксический анализ.

- Парсинг – процесс структурирования линейной репрезентации в соответствии с заданной грамматикой
- Линейной репрезентацией предложения естественного языка называется цепочка элементов, где каждый элемент является минимальной синтаксической единицей

Лингвистический парсер (1)

- ПО для разбора **линейной** последовательности лексем (слов) языка исходного текста во **внутреннее представление** смысла данного П.
- Многоуровневый анализ П. на ЕЯ:

1. Морфологический анализатор

   in: морфологические словари + текст

   out: POS, морфологические признаки

Гаршина В. В., Богоявленская Ю. А. Разработка лингвистического парсера русского языка. // Вестник ВГУ, серия: Системный анализ и информационные технологии, 2012, No 2. http://www.vestnik.vsu.ru/pdf/analiz/2012/02/2012-02-29.pdf

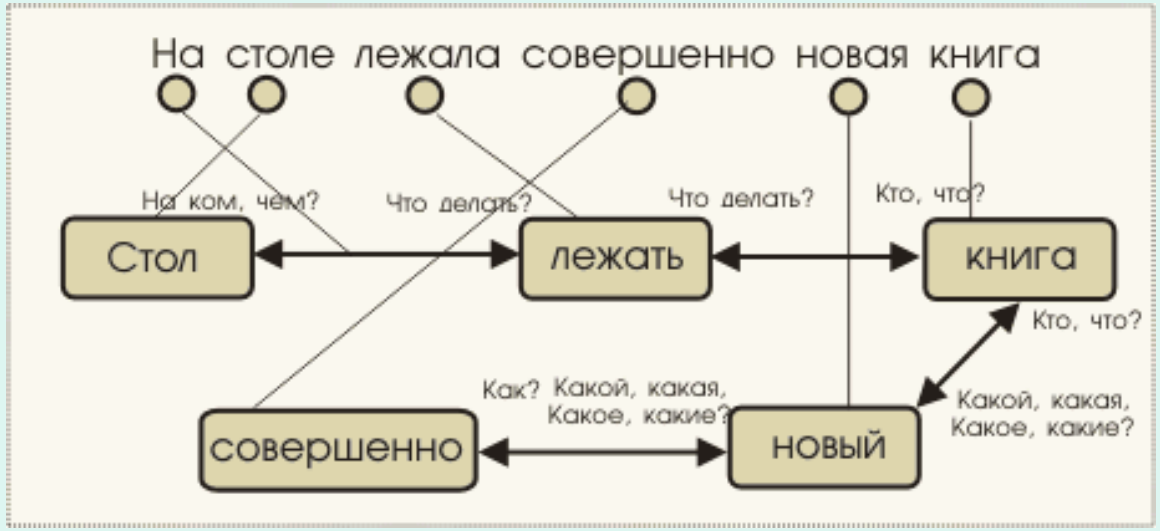
Лингвистический парсер (2)

2. Синтаксичекий анализатор

   out: дерево зависимостей:

   узел: лексема + POS + грамматические хар-ки

   дуга: отношение (подчинения)

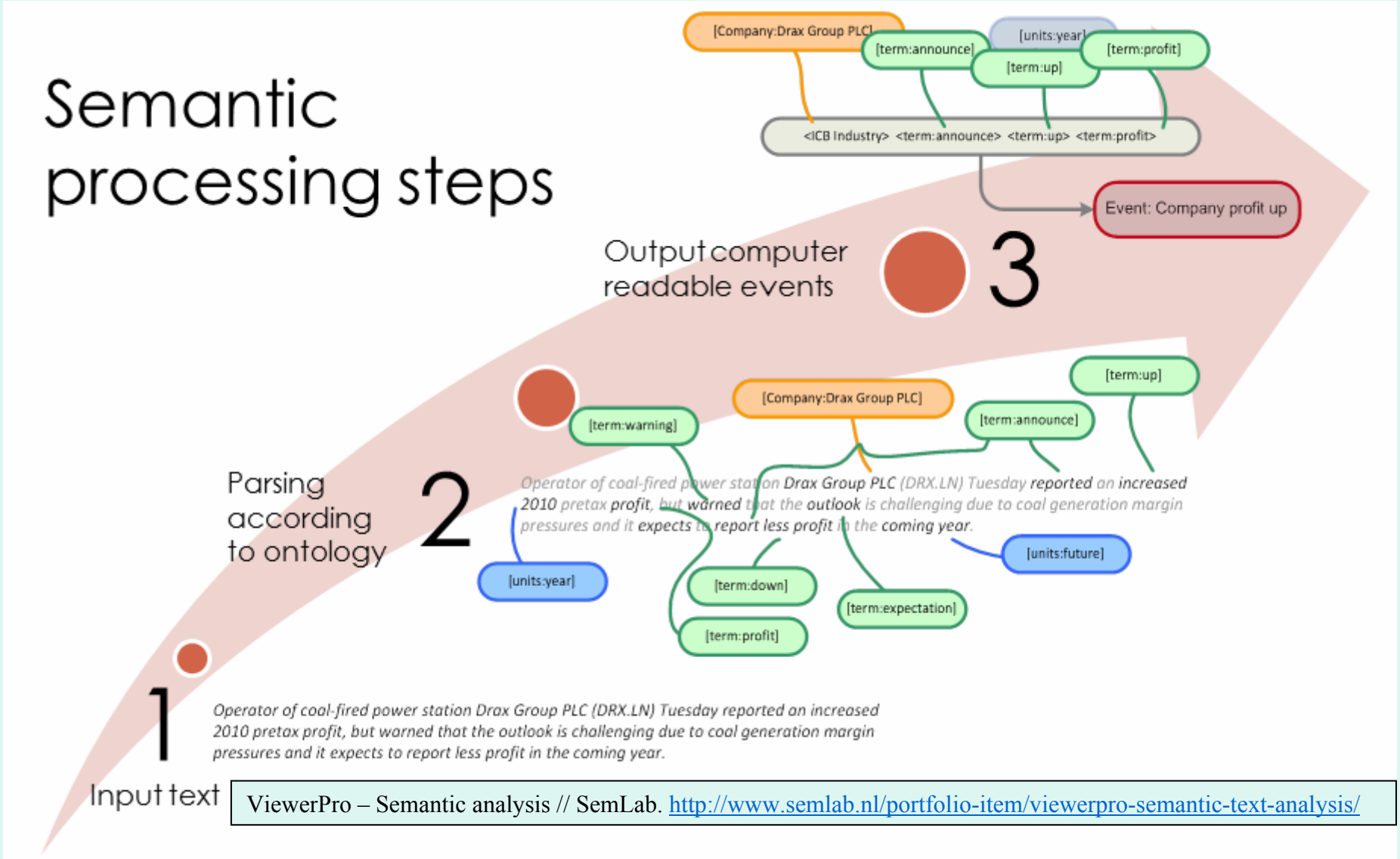


Селезнев К. Технология клиент-сервер // «Открытые Системы», № 12, 2003 http://linter.ru/ru/press-center/detail/27/1554/

3. Семантический анализатор —

in: онтология, предметный словарь, тезаурус

out: дерево зависимостей



ViewerPro – Semantic analysis // SemLab. http://www.semlab.nl/portfolio-item/viewerpro-semantic-text-analysis/

5 языковых средств синтаксического анализа

- Словоизменительные морфологические средства
- *w1 зависит от w2 по C, если граммема g категории C, характеризующая w1 выбирается в зависимости от слова w2*
- *В русском языке к словоизменительным категориям относятся категории падежа и числа существительного; категория падежа числительного; категории рода, числа, падежа и степени сравнения прилагательного; категории лица, числа, времени, наклонения и рода (в прош. вр. и сослагат. накл.) глагола; категория степени сравнения наречия*
- *пример: нов**ые** книг**и***

5 языковых средств синтаксического анализа (2)

- Селективные признаки
  - Частеречные признаки (например, er в конце слова в англ. языке)
  - Одушевленность
- Служебные слова (*в*, *к*, *при, и*, *а*, *или, бы*, *же*, *уж, в течение; несмотря на то что; пускай, давай*)
- Знаки препинания
- Порядок слов

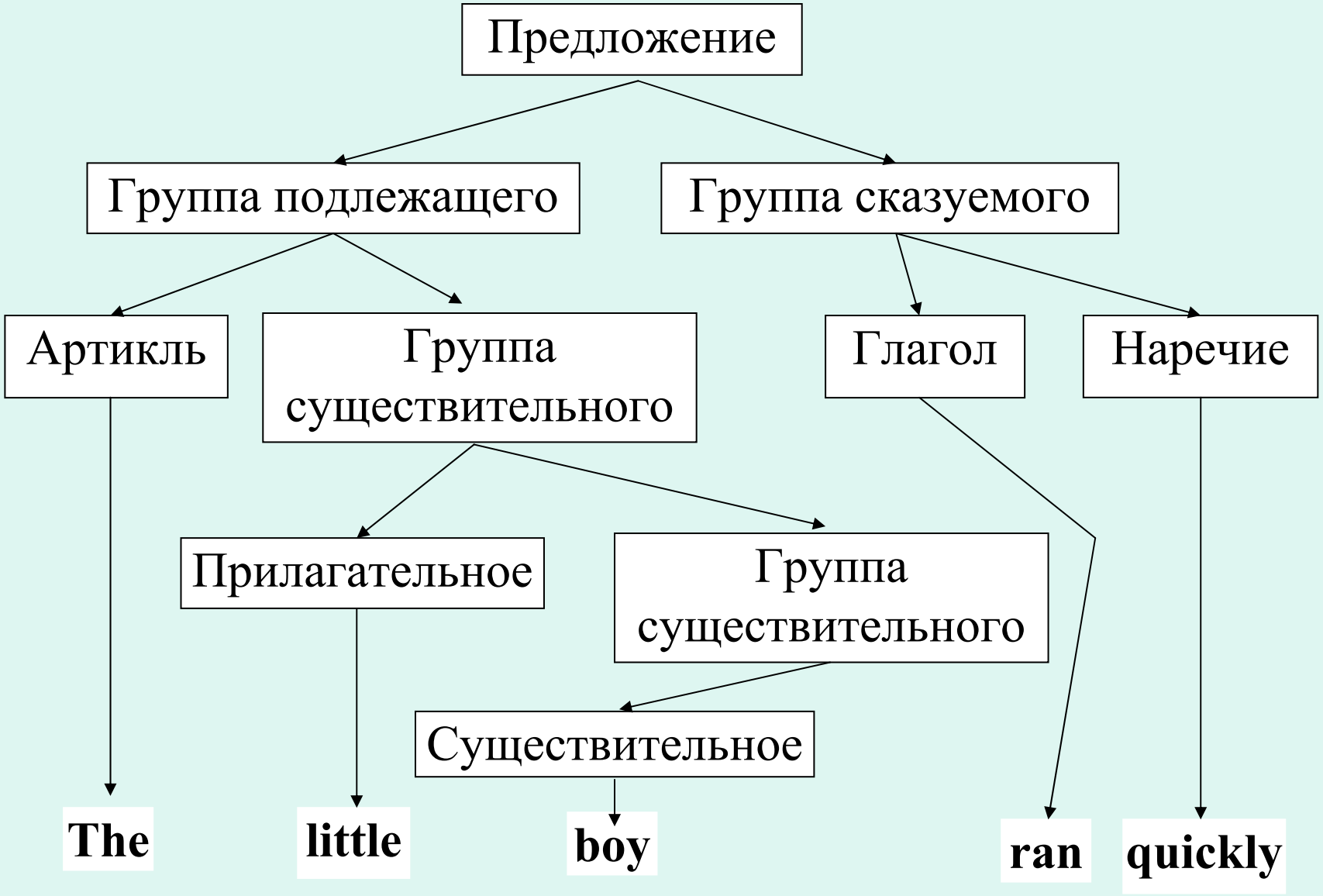
Формализмы

- Контекстно-свободные грамматики (грамматика составляющих)
- Head-driven phrase structure grammar (HPSG)
- Грамматика зависимостей
- Link Grammar
- LR-грамматики

Грамматический разбор

При разборе мы имеем дело с *грамматическими категориями*: ‘предложение’, ‘группа существительного’, ‘группа сказуемого’, ‘существительное’, ‘глагол’, ‘наречие’ и т. д. и пользуемся собственно *словами*, составляющими разбираемое предложение.

Например, структуру английского предложения: “The little boy ran quickly” можно изобразить в виде диаграммы.



Синтаксическая структура предложения

Правила грамматики

Грамматический разбор предложений подразумевает использование *правил* некоторой грамматики. Мы их будем представлять в следующей форме (приведены не все правила грамматики):

< предложение > → < группа подлежащего > < группа сказуемого >

< группа подлежащего > → < артикль > < группа существительного >

< группа существительного > → < существительное >

< группа сказуемого > → < глагол > < наречие >

< артикль > → The

< прилагательное > → little

< существительное > → boy

< глагол > → ran

< наречие > → quickly

Механизм порождения

Здесь стрелочка → отделяет *левую часть правила* от *правой*, а грамматические термины заключены в *металингвистические скобки* < и > для того, чтобы отличать их от слов, составляющих разбираемое предложение.

По этим правилам можно не только *проверять грамматическую правильность предложений*, но также *порождать грамматически правильные предложения*.

Механизм порождения

Начиная с цепочки, включающей только грамматический термин, являющимся *главным* (< предложение >), каждый грамматический термин, входящий в текущую цепочку, замещается правой частью того правила, которое содержит его в левой части. Когда в результате таких замен в текущей цепочке не останется ни одного термина грамматики, а только слова языка, мы получаем *грамматически правильное предложение* языка.

Грамматика.
Язык, порождаемый грамматикой

Ранее речь шла о конкретной грамматике. В ней имеются два словаря:

1) *нетерминалы* — грамматические термины

<предложение>, <группа подлежащего >, …;

2) *терминалы* — слова, составляющие предложения языка

The, little, boy, ran, quickly;

Грамматики

3) *правила*, *левые* и *правые* части которых состоят из нетерминалов и терминалов;

<предложение> → < группа подлежащего >
< группа сказуемого >

< артикль > → The, ...

4) *начальный нетерминал* — главный грамматический термин; из него выводятся те цепочки терминалов, которые считаются предложениями языка

<предложение>

Основные виды грамматик

- Контекстно-свободная грамматика – у которой в левой части правил содержится только один нетерминал

  A -> a, b, c.

- Контекстно-зависимая грамматика – у которой в левой части правил может содержаться помимо нетерминала и терминалы

  Ad -> a, b, c

- Регулярная грамматика – у которой правая часть каждого правила начинается с терминала

  A -> aB

Виды синтаксического анализа (грамматического разбора)

- Сверху вниз
  - Программа пытается породить, начиная с главного правила (описывающего структуру предложения) разбираемое предложение (последовательность терминалов)
- Снизу вверх
  - Программа пытается на основе текущего слова в предложении (и может быть следующих слов) распознать нетерминалы и в конце концов предложение в целом

Технологии анализа ЕЯ. Синтаксический анализ. Свободно-контекстные грамматики.

Недостатки:

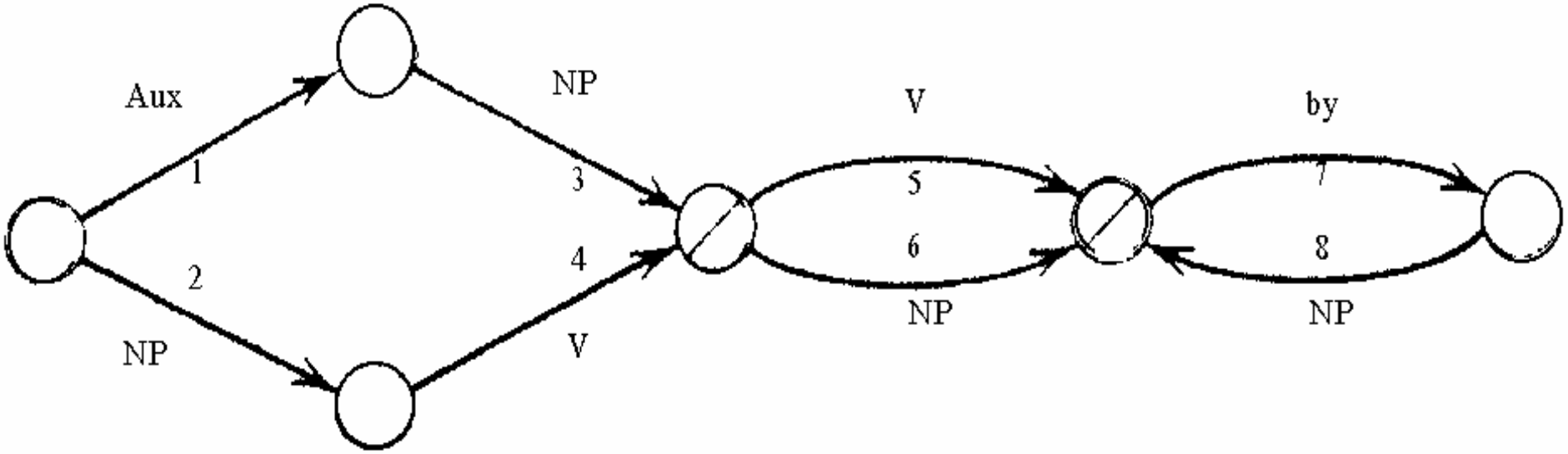
- отсутствие запрета на грамматически неправильные фразы, где, например, подлежащее не согласовано со сказуемым в числе,
- Разрастание грамматики для всех вариантов разбора, в том числе, грамматически неправильных фраз,
- Следствие – непригодны для анализа ЕЯ.

Технологии анализа ЕЯ. Синтаксический анализ. Трансформационные (генеративные) грамматики.

- Автор – Хомский,
- Для порождения грамматически правильных предложений,
- Центральная идея трансформационной теории состоит в том, что поверхностные формы любого языка - его предложения - являются результатом взаимодействия между несколькими модульными подсистемами
- Трансформационные правила предназначены для описания систематических отношений в предложении, как то:
  - отличия между активным и пассивным предложением
  - глобальные отношения в предложении (например, связь между what и eat в предложении "What will John eat")
  - неоднозначности, причиной которых является одна и та же форма предложения, выведенная из двух различных базовых предложений (например, в предложении "They are flying planes" flying можно рассматривать и как прилагательное и как основной глагол)
- Непригодны для анализа ЕЯ.

Технологии анализа ЕЯ. Синтаксический анализ. Расширенная сеть переходов.

- Авторы – Бобров, Фрейзер и Вудс,
- Продолжение свободно-контекстных грамматик.
- *The rabbit nibbles the carrot (Кролик грызет морковь)*.



Технологии анализа ЕЯ. Синтаксический анализ. Расширенная сеть переходов.

Недостатки:

- немодульность;
- сложность при модификации, вызывающая непредвиденные побочные эффекты;
- хрупкость (когда единственная неграмматичность в предложении делает невозможным дальнейший правильный анализ);
- неэффективность при переборе с возвратами, т.к. ошибки на промежуточных стадиях анализа не сохраняются;
- неэффективность с точки зрения смысла, когда с помощью полученного синтаксического представления оказывается невозможным создать правильное семантическое представление.

Технологии анализа ЕЯ. Семантические грамматики.

- Вместо синтаксических категорий в грамматике используются семантические,
- Используются в узких предметных областях,
- Недостатки:
  - для каждой предметной области нужна своя грамматика,
  - Имеют тенденцию быстро увеличиваться в размерах.

S → <present> the <attribute> of <ship>
<present> → what is | [can you] tell me
<ship> → the <shipname> | <classname> class ship

Can you tell me the class of the Enterprise? (*Enterprise* - название корабля).

Технологии анализа ЕЯ. Подбор шаблона.

Диалог программы Eliza, имитирующей психиатра:

X всегда Y:	Вы можете привести какой-нибудь пример?
	Когда?
	Действительно, всегда?
X Вы Y меня:	Почему Вы думаете, что я Y Вас?
	Вам нравится думать, что я Y Вас, не правда ли?
	Представьте, что я Y Вас. Что Вы об этом думаете?

В результате *Элиза* способна вести такой диалог:

Пац.:	Люди **всегда** пристально смотрят на меня.	подбор по 1-му шаблону с учетом *всегда*
Эл.:	Вы можете привести какой-нибудь пример?	первый ответ по 1-му шаблону
Пац.:	Например, сейчас ты **смотришь** на меня.	подбор по 2-му шаблону, Y=*смотреть*
Эл.:	Почему Вы думаете, что я смотрю на Вас?	первый ответ по 2-му шаблону
Пац.:	Потому что ты **ненавидишь** меня.	подбор по 2-му шаблону, Y=*ненавидеть*
Эл.:	Вам нравится думать, что я ненавижу Вас, не правда ли? и т.д.	второй ответ по 2-му шаблону

Язык шаблонов AIML – язык для Alice-подобных диалоговых систем

- AIML – язык разметки
- Набор категорий <Category>
- Категория состоит из
  - Паттернов <pattern>
  - Шаблонов <template>
- При анализе предложения ЕЯ интерпретатор AIML ищет подходящий паттерн, который сопоставим с предложением, а в ответ выдает один из соответствующих шаблонов
- Имеется возможность запоминать слова из предложения и использовать их в виде переменных в шаблонах и при проверке условий

Технологии анализа ЕЯ. Падежные фреймы.

- Падежный фрейм состоит из заголовка и набора ролей (падежей), связанных определенным образом с заголовком

[Заголовочный глагол
[падежный фрейм
агент: <активный агент, совершающий действие>
объект: <объект, над которым совершается действие>
инструмент: <инструмент, используемый при совершении действия>
реципиент: <получатель действия - часто косвенное дополнение>
направление: <цель (обычно физического) действия>
место: <место, где совершается действие>
бенефициант: <сущность, в интересах которой совершается действие>
коагент: <второй агент, помогающий совершать действие>
]]

Технологии анализа ЕЯ. Падежные фреймы.

Например,
для фразы *Иван дал мяч Кате* падежный фрейм выглядит так:
[Давать
[падежный фрейм
агент: Иван
объект: мяч
реципиент: Катя]
[грам
время: прош
залог: акт]
]

Технологии анализа ЕЯ. Падежные фреймы.

Анализ текста с помощью падежных фреймов состоит из следующих шагов:

- Используя существующие фреймы, подобрать подходящий для заголовка. Если такого нет, текст не может быть проанализирован.
- Вернуть в систему подходящий фрейм с соответствующим заголовком-глаголом.
- Попытаться провести анализ по всем обязательным падежам. Если один или более обязательных заполнителей падежей не найдены, вернуть в систему код ошибки. Такой случай может означать наличие эллипсиса, неверный выбор фрейма, неверно введенный текст или недостаток грамматики. Следующие шаги используются уже для анализа и исправления таких ситуаций.
- Провести анализ по всем необязательным падежам.
- Если после этого во введенном тексте остались непроанализированные элементы, выдать сообщение об ошибке, связанной с неправильным вводом, недостаточностью данного анализа или необходимостью провести другой, более гибкий анализ.

Технологии анализа ЕЯ. Падежные фреймы.

Преимущества использования падежных фреймов таковы:

- совмещение двух стратегий анализа (сверху вниз и снизу вверх);
- комбинирование синтаксиса и семантики;
- легкая встраиваемость в интеллектуальные системы на очнове фреймов;
- удобство при использовании модульных программ.