Разработка средств автоматического синтаксического анализа как модуля системы понимания текста

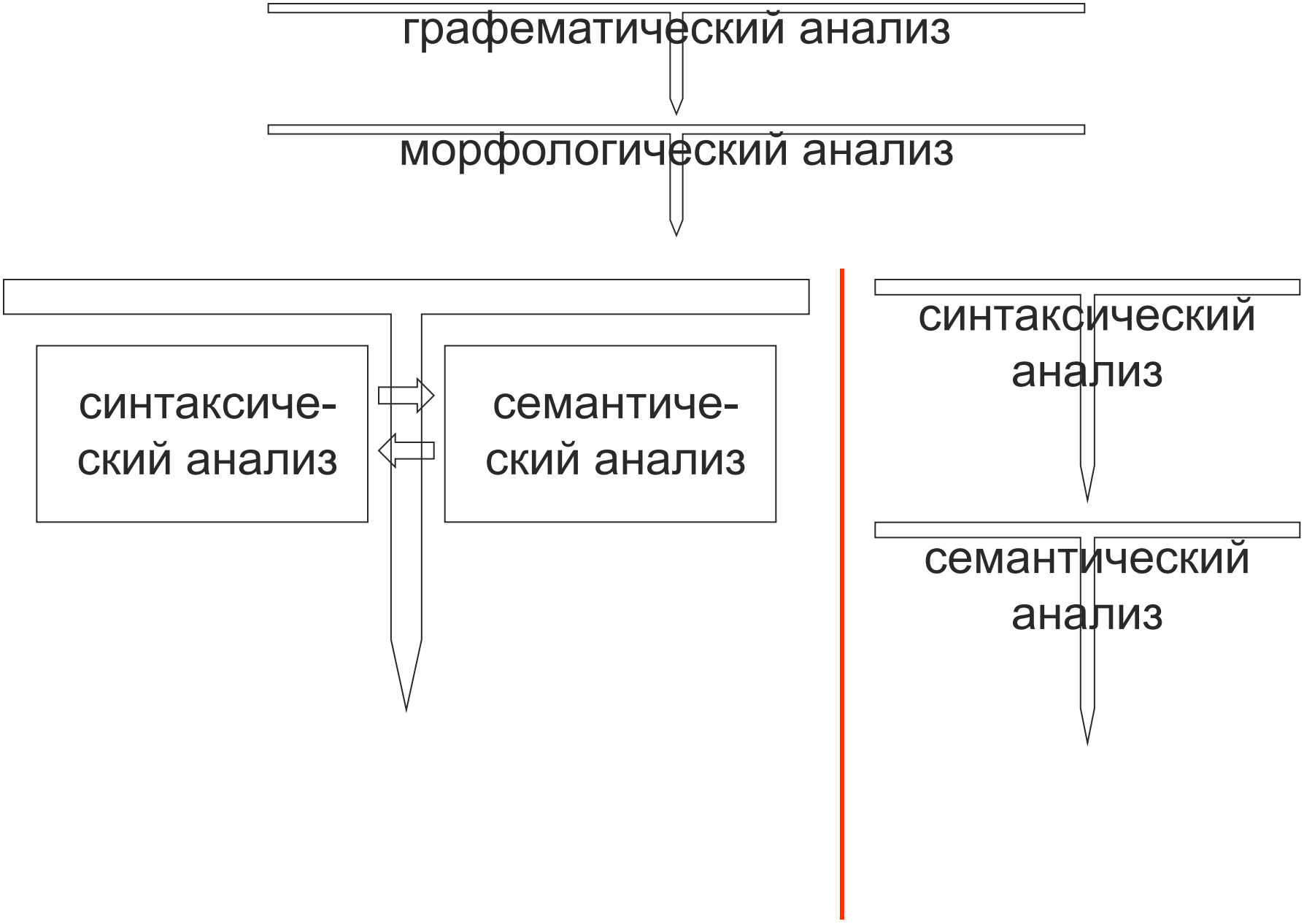
Лахути Д.Г., Баталина А.М., Епифанов М.Е., Кобзарева Т.Ю.

(РГГУ)

26 марта 2009 г.

Что значит для нас понять следующее предложение:

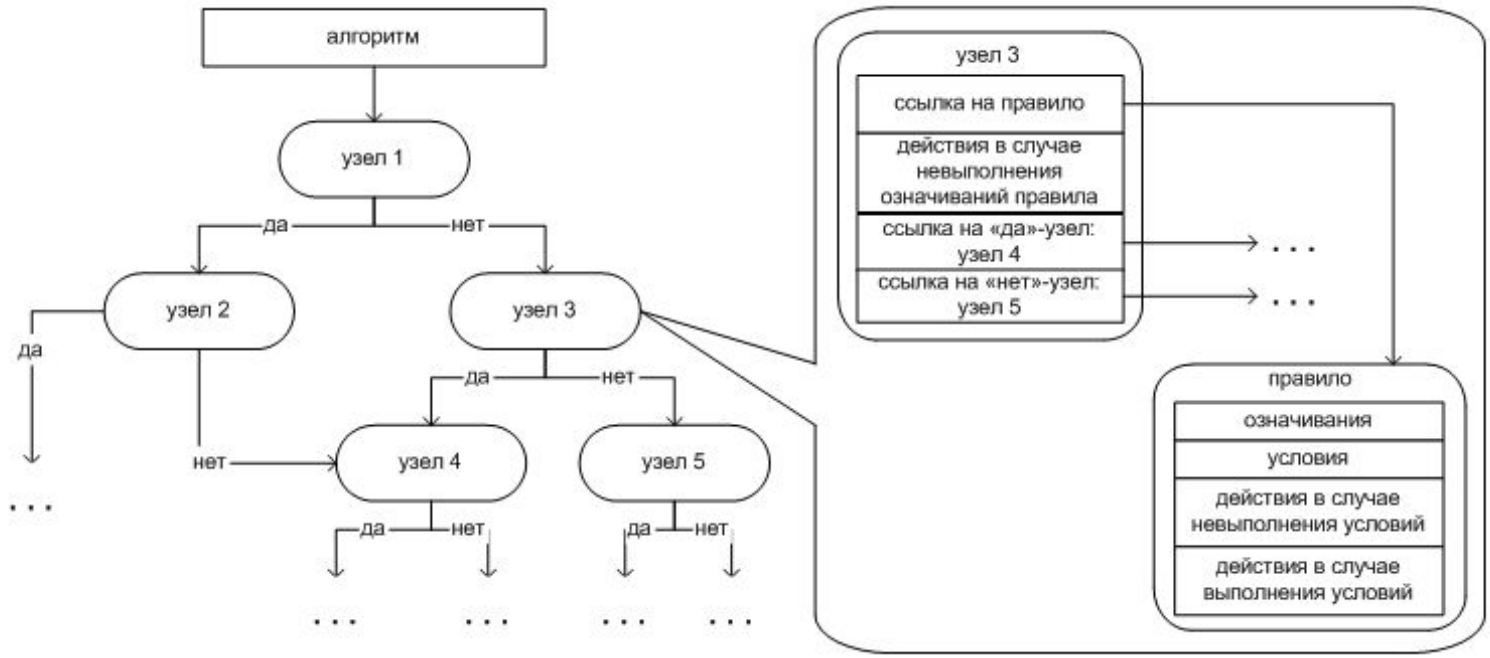
Императрикс Елисавета, о! приехала в Царское Село.


графематический анализ
морфологический анализ
синтаксиче-
ский анализ
семантиче-
ский анализ
синтаксический
анализ
семантический
анализ

Пример синтаксической неоднозначности:

В этом музее были выставлены чучела динозавров, которые все погибли от бомбёжки.

Объектная модель алгоритма в инструментальной среде



Объектная модель анализируемого предложения

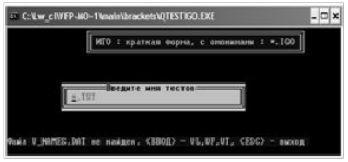
- лексические единицы и их свойства в формате (*имя_свойства значение_свойства*)
- граф синтаксических связей в формате (*имя_синтаксического_хозяина имя_синтаксического_слуги тип_связи*).

Выбор средств реализации среды для экспериментов с алгоритмами ПСА

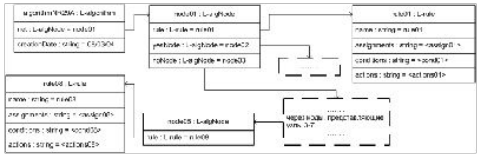
- реализация выполнена на языке Common Lisp с использованием встроенной в него библиотеки CLOS (Common Lisp Object System) в инструментальной среде разработки Corman Common Lisp ® версии 2.5
- в Лиспе имеется возможность в процессе выполнения функций вычислять выражения, записанные в его же синтаксисе - можно строить код новых Лисп-функций во время работы программы и, при желании, вычислять их при каких-либо значениях аргументов
- Лисп – интерпретируемый язык, в рассматриваемой среде алгоритмы рассматриваются как данные по отношению к интерпретатору
- для записи алгоритмов используется своеобразный предметно ориентированный языковой «конструктор» (при составлении описания алгоритма мы или манипулируем – добавляем, переставляем, удаляем – готовыми «кирпичиками» - объектами, представляющими правила, или только меняем содержание правил), что обеспечивает легкость модификации

Инструментальная среда и окружающие ее приложения

Морфолопический анализатор «Скобки» (Информэлектро)



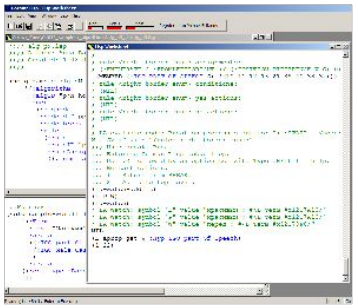
Графический конструктор объектов (Ершова Е.С., Колесников А.А.)



конверсия результата работы в объектную модель инструментальной среды

конверсия графического описания алгоритма по его xml-описанию в объектную модель инструментальной среды

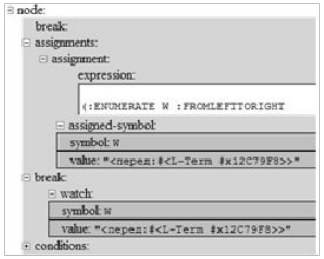
Инструментальная среда для экспериментов с алгоритмами поверхностно-синтаксического анализа



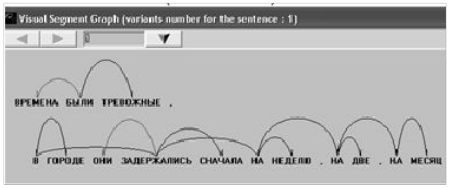
построение xml-описание для общения с приложением

построение xml-описания для общения с приложением

Дерево прохода алгоритма (Г.Ю.Айриян)



Графическое отображение связей, сегментов и характеристик лексических единиц примера (И.М.Ножов)



Входные языки для инструментальной среды

- входной предметно ориентированный лиспообразный язык для описания алгоритма (для трансляции в объектную модель инструментальной среды)
- языки в синтаксисе xml (служат посредниками между инструментальной средой и внешними программами для просмотра результатов)
- графический (используется в графическом редакторе структуры объектов, ориентированном на объектную модель инструментальной среды)

Схема работы в инструментальной среде

Запись лингвистического алгоритма на предметно ориентированном входном языке

Отладка алгоритма на примере (предложении)

Просмотр результата применения алгоритма к примеру

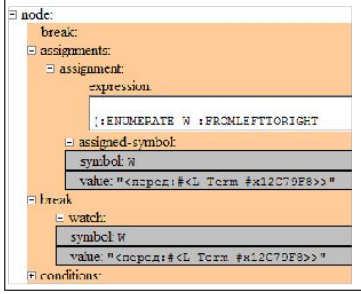
```
'(:algorithm
  :algID "alg_p_ed.izm"
 :net
(:algnode
   :nodeid "n1"
   :rule
   (:rule
    :ruleID "1"
    :ruleAssignments
    ((:enumerate W :fronRightToLeft (equal (:Hyp-IGO-Part-of-Speech W) '(2 22)) )  )
    :ruleConditions
              ((and (not (null (:neighbour W :toLeft 1) ))
                    (or (member (:IGO-Part-of-Speech (:neighbour W :toLeft 1)) '(3 4 31))
                          (equal (:Hyp-IGO-Part-of-Speech (:neighbour W :toLeft 1)) '(28 4))
                          (equal (:Hyp-IGO-Part-of-Speech (:neighbour W :toLeft 1)) '(4 28))
                          )) )
    :ruleYesActions ((:fix-homonym-as W '(2)))
    :ruleNoActions ((if (and (boundp 'W) (equal (:Hyp-IGO-Part-of-Speech W) '(2 22)) )
                        (:fix-homonym-as W '(22))) )

   )
   :node-when-ass-failed :return
   :noNode :repeat
   :yesNode :repeat
   )
)
```

Установка прерываний в нужных частях алгоритма

Блуждание по узлам алгоритма

Выявление неточностей / ошибок в алгоритме

Дерево прохода алгоритма (Г.Ю.Айриян)



Графическое отображение связей, сегментов и характеристик лексических единиц примера (И.М.Ножов)

Функциональность инструментальной среды для экспериментов с алгоритмами поверхностно-синтаксического анализа

- установка прерывания
- пошаговое выполнение алгоритмов
- просмотр на каждом шаге значений переменных, содержания узла и связанного с узлом правила, текущего состояния представления анализируемого предложения
- вычисление пробных вариантов правил и используемых в них форм
- протоколирование выполнения алгоритма
- механизм пакетного тестирования – возможность автоматически применять составленные ранее тесты к объекту-результату вычисления алгоритма на некотором примере
- поддержка т.н. проектов тестирования – файлов, в которых в соответствующем формате записана информация о соответствии тестовых примеров алгоритмам, о различных версиях алгоритма и т.п.

Промоделированы и в основном отлажены/находятся в стадии отладки

- Блоки предсинтаксиса и предсегментации
  - модули предсинтаксиса:
    - стандартные универсальные подпрограммы проверки согласования,
    - алгоритмы постморфологии, корректирующие и дополняющие результаты морфологического анализа,
    - наиболее актуальные алгоритмы снятия омонимии частей речи,
  - часть алгоритмов модуля предсегментации:
    - построение атрибутивных именных групп и предложных групп,
    - построение конструкций с именами собственными, с числами,
    - построение сложных сказуемых,
    - построение синтагм со слугами – обособленными приложениями.

Промоделированы и в основном отлажены/находятся в стадии отладки

- Блок сегментации
  - экспресс-версия сегментационного анализа (не предполагается возможности любых разрывающих вложений так называемых а-сегментов – придаточных предложений, обособленных согласованных определений, деепричастных, предложных, вводных и сравнительных оборотов – в а-сегменты).
  - полная версия сегментационного анализа (рассчитан на сегментацию любых грамматически правильных неэллиптичных предложений литературного письменного языка, не являющихся записью или имитацией устной речи).
- Блок внутрисегментного анализа
  - поиск сказуемого и подлежащего,
  - заполнение словарно заданных валентностей,
  - поиск хозяина предложной группы,
  - поиск хозяев слабоуправляемых именных групп в родительном падеже и наречий.