

Компьютерный анализ естественно-языкового текста

Кафедра информационных систем в
искусстве и гуманитарных науках

СТРУКТУРА КУРСА

1. Введение в дисциплину
2. Автоматический анализ текста на морфологическом уровне
3. Автоматический анализ текста на синтаксическом уровне
4. Семантический компонент в системах автоматического анализа текста

СТРУКТУРА КУРСА

2. Автоматический анализ текста на морфологическом уровне
 1. *Морфологический уровень в ЛИТ*
 2. *Основные понятия морфологии в компьютерной морфологии*
 3. *Основные процедуры компьютерной морфологии*
 4. *Компьютерная морфология русского языка*
 5. *Технологии морфологического анализа*
 6. *«Предсказание» (типизация)*
 7. *Вопросы, смежные с синтаксисом*

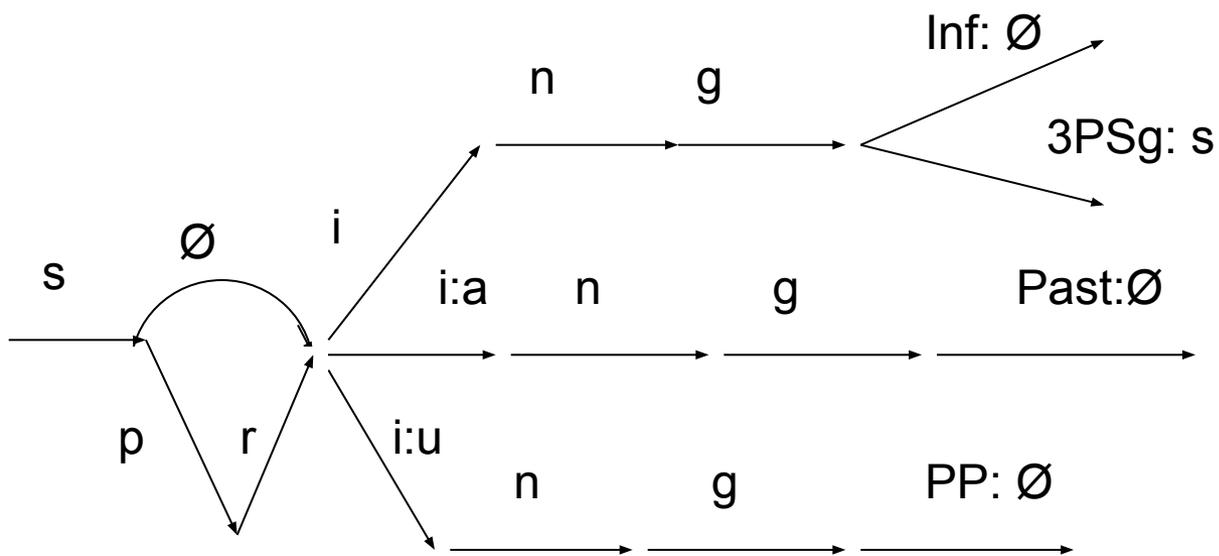
ФОРМАЛЬНО-ЛИНГВИСТИЧЕСКИЙ СМЫСЛ КОНЕЧНОГО ПРЕОБРАЗОВАТЕЛЯ

- Конечный автомат – язык
- Конечный преобразователь – отношение
- Язык: $L \subseteq V^*$
- Отношение: $R \subseteq V_B^* \times V_H^*$

ФРАГМЕНТ АНГЛИЙСКОЙ ГЛАГОЛЬНОЙ СИСТЕМЫ

- sing+Inf:sing;
- sing+3pSg:sings;
- sing+Past:sang;
- sing+PP:sung;
- spring+Inf:spring;
- spring+3pSg:springs;
- spring+Past:sprang;
- spring+PP:sprung;

ФРАГМЕНТ АНГЛИЙСКОЙ ГЛАГОЛЬНОЙ СИСТЕМЫ: КОНЕЧНЫЙ ПРЕОБРАЗОВАТЕЛЬ



ЛЕКСИКОН В ФОРМАТЕ Xerox Tools

Multichar_Symbols +Inf+3pSg +Past +PP

LEXICON Root

sing+Inf:sing # ;

sing+3pSg:sings # ;

sing+Past:sang # ;

sing+PP:sung # ;

spring+Inf:spring # ;

spring+3pSg:springs # ;

spring+Past:sprang # ;

spring+PP:sprung # ;

ЛИСТИНГ РАБОТЫ С XFST (1)

>xfst

ЛИСТИНГ РАБОТЫ С XFST (2)

>xfst

Copyright © Palo Alto Research Center 2001-2007
Xerox Finite-State Tool, version 2.6.2

Type "help" to list all commands available or "help help" for further help.

xfst[0]:

ЛИСТИНГ РАБОТЫ С XFST (3)

>xfst

Copyright © Palo Alto Research Center 2001-2007
Xerox Finite-State Tool, version 2.6.2

Type "help" to list all commands available or "help help" for further help.

xfst[0]: read lexc ex3c.txt

ЛИСТИНГ РАБОТЫ С XFST (4)

Copyright © Palo Alto Research Center 2001-2007
Xerox Finite-State Tool, version 2.6.2

Type "help" to list all commands available or "help help" for further help.

```
xfst[0]: read lexc ex3c.txt
```

```
Reading ISO-8859-1 text from 's-lex.txt'
```

```
Root...2
```

```
Building lexicon...Minimizing...Done!
```

```
1.5 Kb. 14 states, 19 arcs, 8 paths. Label Map: Default
```

```
Closing 's-lex.txt'
```

```
xfst[1]:
```

ЛИСТИНГ РАБОТЫ С XFST (5)

Copyright © Palo Alto Research Center 2001-2007
Xerox Finite-State Tool, version 2.6.2

Type "help" to list all commands available or "help help" for further help.

```
xfst[0]: read lexc ex3c.txt
```

```
Reading ISO-8859-1 text from 's-lex.txt'
```

```
Root...2
```

```
Building lexicon...Minimizing...Done!
```

```
1.5 Kb. 14 states, 19 arcs, 8 paths. Label Map: Default
```

```
Closing 's-lex.txt'
```

```
xfst[1]: print words
```

ЛИСТИНГ РАБОТЫ С XFST (6)

```
Closing 's-lex.txt'  
xfst[1]: print words  
sing+3pSg:s  
sing+Inf:0  
si:ang+Past:0  
si:ung+PP:0  
spring+3pSg:s  
spring+Inf:0  
spri:ang+Past:0  
spri:ung+PP:0  
xfst[1]:
```

ЛИСТИНГ РАБОТЫ С XFST (7)

```
Closing 's-lex.txt'  
xfst[1]: print words  
sing+3pSg:s  
sing+Inf:0  
si:ang+Past:0  
si:ung+PP:0  
spring+3pSg:s  
spring+Inf:0  
spri:ang+Past:0  
spri:ung+PP:0  
xfst[1]: print net
```

ЛИСТИНГ РАБОТЫ С XFST (8)

xfst[1]: print net

Sigma: a g i n p r s u +Inf +PP +Past +3pSg

Size: 12, Label Map: Default

Net:

Flags: deterministic, pruned, minimized, epsilon_free,
loop_free

Arity: 2

s0: s -> s1.

s1: i -> s2, p -> s3, i:a -> s4, i:u -> s5.

s2: n -> s6.

s3: r -> s7.

s4: n -> s8.

ЛИСТИНГ РАБОТЫ С XFST (9)

s0: s -> s1.

s1: i -> s2, p -> s3, i:a -> s4, i:u -> s5.

s2: n -> s6.

s3: r -> s7.

s4: n -> s8.

s5: n -> s9.

s6: g -> s10.

s7: i -> s2, i:a -> s4, i:u -> s5.

s8: g -> s11.

s9: g -> s12.

s10: +Inf:0 -> fs13, +3pSg:s -> fs13.

ЛИСТИНГ РАБОТЫ С XFST (10)

s4: n -> s8.

s5: n -> s9.

s6: g -> s10.

s7: i -> s2, i:a -> s4, i:u -> s5.

s8: g -> s11.

s9: g -> s12.

s10: +Inf:0 -> fs13, +3pSg:s -> fs13.

s11: +Past:0 -> fs13.

s12: +PP:0 -> fs13.

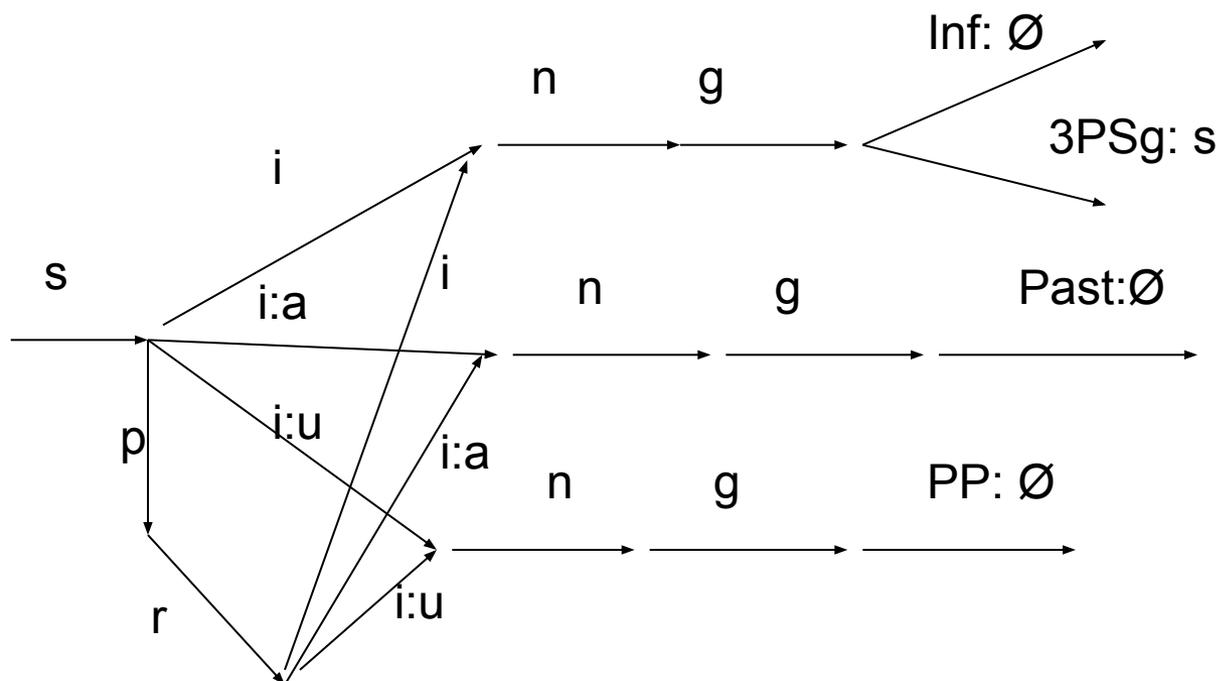
fs13: (no arcs)

xfst[1]:

ФРАГМЕНТ АНГЛ. ГЛАГОЛЬНОЙ СИСТЕМЫ: КОНЕЧ. ПРЕОБРАЗОВАТЕЛЬ (версия Xerox)

s0: s -> s1.
s1: i -> s2, p -> s3, i:a -> s4, i:u -> s5.
s2: n -> s6.
s3: r -> s7.
s4: n -> s8.
s5: n -> s9.
s6: g -> s10.
s7: i -> s2, i:a -> s4, i:u -> s5.
s8: g -> s11.
s9: g -> s12.
s10: +Inf:0 -> fs13, +3pSg:s -> fs13.
s11: +Past:0 -> fs13.
s12: +PP:0 -> fs13.
fs13: (no arcs)

ФРАГМЕНТ АНГЛ. ГЛАГОЛЬНОЙ СИСТЕМЫ: КОНЕЧ. ПРЕОБРАЗОВАТЕЛЬ (версия Xerox)



КОНЕЧНЫЙ ПРЕОБРАЗОВАТЕЛЬ КАК СРЕДСТВО МОРФ. АНАЛИЗА И СИНТЕЗА

xfst[1]: up spring

spring+Inf

xfst[1]: up sprang

spring+Past

xfst[1]: up sung

sing+PP

xfst[1]: down sing+3pSg

sings

xfst[1]:

ЛЕКСИКОН В ФОРМАТЕ Xerox Tools

Multichar_Symbols +Inf+3pSg +Past +PP

LEXICON Root

sing+Inf:sing # ;

sing+3pSg:sings # ;

sing+Past:sang # ;

sing+PP:sung # ;

spring+Inf:spring # ;

spring+3pSg:springs # ;

spring+Past:sprang # ;

spring+PP:sprung # ;

РАСШИРЕНИЕ ЛЕКСИКОНА

LEXICON Root

sing+Inf:sing # ;

sing+3pSg:sings # ;

spring+Inf:spring# ;

spring+3pSg:springs # ;

sprint+Inf:sprint #;

sprint+3pSg:sprints #;

sprout+Inf:sprout #;

sprout+3pSg:sprouts #

spruce+Inf:spruce #;

spruce+3pSg:spruces #

spud+Inf:spud #;

РАСШИРЕНИЕ ЛЕКСИКОНА на ЯЗЫКЕ LEXC

LEXICON Root

sing Ending ;

spring Ending ;

sprint Ending ;

sprout Ending ;

spruce Ending ;

spud Ending ;

LEXICON Ending

+Inf:0 #;

+3pSg:s #;

РЕЗУЛЬТАТ РАСШИРЕНИЯ ЛЕКСИКОНА

xfst[1]: print words

sing+Inf:0

sing+3pSg:s

sprint+Inf:0

sprint+3pSg:s

spring+Inf:0

spring+3pSg:s

sprout+Inf:0

sprout+3pSg:s

spruce+Inf:0

spruce+3pSg:s

spud+Inf:0

spud+3pSg:s

ОБЩИЕ СВЕДЕНИЯ О ЗАПИСИ ЛЕКСИКОНОВ НА ЯЗЫКЕ LEXC (1)

Объявление комплексных символов (Multichar_Symbols) –
факультативно

Определение классов (Definitions) – *факультативно*

LEXICON Root

LEXICON A

LEXICON B

END - *факультативно*

ОБЩИЕ СВЕДЕНИЯ О ЗАПИСИ ЛЕКСИКОНОВ НА ЯЗЫКЕ LEXS (2)

Все записи во всех лексиконах обязательно представляют собой форму (терминальную цепочку, возможно, нулевую) и класс продолжений (имя одного из последующих лексиконов)

Каждая запись в лексиконе оканчивается на ";"

Комментарии вводятся знаком "!"

Для буквальной интерпретации следующего символа используется знак "%" (перед ;#;!0<>)

ТИПЫ СЛОВАРНЫХ СТАТЕЙ В ЛЕКСИКОНАХ LEXC

LEXICON Root

go# ;

go:went #;

< d o:i 0:d > # ;

<a b c* (d) e+ > # ;

РЕГУЛЯРНЫЕ ВЫРАЖЕНИЯ

Обобщающий термин для средств записи регулярных языков и регулярных отношений

- Конечный автомат соответствует регулярному языку
- Конечный преобразователь соответствует регулярному отношению

ФОРМАЛЬНО-МАТЕМАТИЧЕСКИЙ СМЫСЛ РЕГУЛЯРНОГО ЯЗЫКА

- Регулярный язык - формальный язык, специфика которого заключается в способе определения:
определяется по образцу алгебраического исчисления, т.е. через исходный алфавит и набор операций, которые могут быть применены к символам этого алфавита, образуя цепочки определяемого языка.
- Множество всех возможных регулярных языков над заданным алфавитом - результат всех возможных применений операций определенного класса

РАЗГРАНИЧЕНИЕ ПОНЯТИЙ (1)

- Символ a
- Цепочка “ a ”
- Язык $\{“a”\}$

- Язык $\{“a”\}$
- Регулярное выражение a
- Конечный автомат (задается диаграммой или таблицей)
- Автоматная грамматика (задается набором правил)

РАЗГРАНИЧЕНИЕ ПОНЯТИЙ (2)

- Регулярным выражением **обозначается** язык
- Регулярное выражение **компилируется** в виде конечного автомата
- Язык **представляется** в виде конечного автомата

ОСНОВНЫЕ ОПЕРАЦИИ РЕГУЛЯРНЫХ ЯЗЫКОВ

- конкатенация ($a b$ или $\{ab\}$)
- итерация ($*$ и $+$)
- факультативность (заключение в круглые скобки)
- объединение ($|$)
- отрицание/дополнение (\sim) и термовое отрицание/дополнение (\setminus)
- пересечение ($\&$)

ФОРМАЛЬНО-МАТЕМАТИЧЕСКИЙ СМЫСЛ КОНЕЧНОГО ПРЕОБРАЗОВАТЕЛЯ

- Конечный автомат – регулярный язык
- Конечный преобразователь –
регулярное отношение
- Регулярное отношение: Результат
объединения произведений регулярных
языков

РАЗГРАНИЧЕНИЕ ПОНЯТИЙ

- Символ a
- Цепочка “ a ”
- Язык $\{“a”\}$

- Пара символов $a:a$
- Пара цепочек “ $a:a$ ”
- Отношение $\{“a:a”\}$