
ЛЕКСИКО-СИНТАКСИЧЕСКИЕ ШАБЛОНЫ В ЗАДАЧАХ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТА

Большакова Е.И., Баева Н.В., Бордаченкова Е.А.,
Васильева Н.Э., Морозов С.С.

МГУ им. М.В. Ломоносова Факультет ВМиК

bolsh@cs.msu.su

СОДЕРЖАНИЕ ДОКЛАДА

1. Задача формального описания лексических и морфосинтаксических особенностей текстовых единиц.
 2. Результаты сравнительного анализа средств описания (НКРЯ, Alex, RCO).
 3. Концепция лексико-синтаксического шаблона языковых конструкций.
 4. Основные возможности языка записи лексико-синтаксических шаблонов (далее LSPL).
-

ЗАДАЧА ОПИСАНИЯ ЯЗЫКОВЫХ КОНСТРУКЦИЙ

Изучение терминологических и дискурсивных особенностей

НТ прозы



Потребность формализовать характерные конструкции
(*Под T будем понимать D, Далее докажем P, Допустим, что S*)



Определение множества лексем, грамматических форм,
синтаксических условий



Фиксирование в виде декларативной структуры –
лексико-синтаксического шаблона языковой конструкции

NG_{ACC} [«мы»] «будем называть» T_{INS}

СРЕДСТВА ОПИСАНИЯ ЕДИНИЦ ТЕКСТА ДЛЯ ПОИСКА ФРАГМЕНТОВ В ТЕКСТАХ НА РУССКОМ ЯЗЫКЕ

Единицы текста: словоформы, последовательности слов, словосочетания

■ Национальный корпус русского языка (НКРЯ)

- ❑ ограниченность корпуса; лингвистическая разметка
- ❑ поиск последовательности слов по их грамматическим и лексико-семантическим характеристикам

■ Система Alex

- ❑ лексические шаблоны для узкоспециализированных текстов
- ❑ средства описания словосочетаний, без указания грамматических признаков

■ RCO Pattern Extractor/система GATE

- ❑ правила и шаблоны для извлечения из текста специфических объектов
- ❑ формальный язык в стиле ЯП (атрибутно-объектная модель текста)

СРАВНЕНИЕ ЯЗЫКОВЫХ СРЕДСТВ: ОПИСАНИЕ ЛЕКСИКО-ГРАММАТИЧЕСКИХ ОСОБЕННОСТЕЙ

Лексико-графические единицы

Конкретная словоформа	<i>Позволяют все средства</i>
Произвольная символьная строка из буквенных и небуквенных символов	<i>НКРЯ не производит поиск строк со знаками препинания</i>
Произвольная словоформа в рамках лексемы	<i>Недоступно в Alex без описания шаблона всех словоформ</i>

Морфо-синтаксические условия

Морфологические характеристики (часть речи, падеж, число, время)	<i>Есть в НКРЯ и RCO</i>
Грамматическое согласование нескольких единиц	<i>Нельзя непосредственно записать ни в одной из систем</i>

СРАВНЕНИЕ ЯЗЫКОВЫХ СРЕДСТВ: ДОПОЛНИТЕЛЬНЫЕ ВОЗМОЖНОСТИ

Логические операции

Комбинирование лексических условий	<i>Есть во всех системах</i>
Комбинирование морфосинтаксических условий	<i>Есть в НКРЯ и RCO</i>

Запись конструкций

Альтернативы и повторения	<i>Отсутствуют в НКРЯ</i>
Именованые конструкции	<i>Возможно в Alex и RCO</i>

ЛЕКСИКО-СИНТАКСИЧЕСКИЙ ШАБЛОН

Разработка формального языка для:

- записи специфических языковых конструкций для их представления в системе автоматической обработки НТТ;
- записи запросов на поиск конструкций для системы поддержки лингвистических исследований.

Лексико-синтаксический шаблон – структурный образец языковой конструкции, отображающий ее *лексические* и *поверхностно-синтаксические* свойства.

Принцип отбора выразительных средств:

- гибкая и интуитивно понятная запись основных лексических и поверхностно-синтаксических свойств конструкций.
-

ЯЗЫК LSPL-ШАБЛОНОВ: ОСНОВНЫЕ ВОЗМОЖНОСТИ

Элемент-слово включает:

- часть речи (A, N, V, Pa и т.д.) – A
- индекс – A1 A2 N
- лексема (< >) – A<важный>
- уточнение грамматических характеристик (имя=значение) – A<важный; case=nom, gen=fem>

Грамматическое согласование элементов шаблона:

A<тяжелый> N <A.gen=N.gen, A.num=N.num, A.case=N.case>

A<тяжелый> N <A=N>

Слово *тяжелый* и следующее за ним существительное согласованы в роде, числе и падеже: *тяжелым вечером, тяжелых камней, тяжелое тело*

ЯЗЫК LSPL-ШАБЛОНОВ: ДОПОЛНИТЕЛЬНЫЕ ВОЗМОЖНОСТИ

- $AP = A(A) | Pa(Pa)$
- $AS = \{AP\} N \langle \text{стол, } c = \text{nom} \rangle [\text{“в”}] \langle AP = N \rangle (N)$

Элемент-слово

Альтернативы |

Имя шаблона

Повторение { }

Экземпляр шаблона

Оptionальное вхождение []

Условия согласования

Параметры шаблона

ДОПОЛНИТЕЛЬНЫЕ ПРИМЕРЫ

- Однородные члены в виде именных групп:

SNG = AN1 {“,” AN2}<1> [“и” AN3] <AN1.c=AN2.c=AN3.c> (AN1)

Дама сдавала в багаж диван, чемодан, саквояж, картину, корзину, картонку и маленькую собачонку

- Шаблон типичной для деловой и НТ прозы конструкции:

NP = AN1 {AN2<case=gen>} (AN1)

- Характерная конструкция определения новых терминов:

DT = NP1<c=acc> ["мы"] "назовем" NP2<c=ins> <NP1.n = NP2.n>

*Указанную операцию **назовем** операцией поиска примеров*

ЯЗЫК LSPL-ШАБЛОНОВ: СРАВНИТЕЛЬНЫЙ ПРИМЕР

Прилагательное и существительное
в именительном падеже единственного числа

- ЯЗЫК LSPL:

A<c=nom, n=sign> N<c=nom, n=sign>

- ЯЗЫК RCO Pattern Extractor:

{Morph.SpeechPart="Noun", Morph.Case="Nominative",
Morph.Number="Singular"}

{Morph.SpeechPart="Adjective", Morph.Case="Nominative",
Morph.Number="Singular"}

ЗАКЛЮЧЕНИЕ

- Разработана первая версия программного модуля для поиска в тексте фрагментов, соответствующих заданному LSPL-шаблону.
- Изучаются возможности развития языка LSPL:
 - усиление его выразительности:
 - логическое комбинирование условий;
 - грамматическое управление;
 - введение операций над фрагментами:
 - подсчет статистики;
 - извлечение составных конструкций.

СПАСИБО ЗА ВНИМАНИЕ!