Сравнительный анализ языков описания лексико-синтаксических шаблонов

Гурова А., Иксанова А.

План

- JAPE
- LSPL (Большакова Е.И.)
- LSPL (Рабчевский Е.А,)
- Томита
- Сравнение и выводы

JAPE

- Используется при работе с GATE
- Тексты на английском и других языках
- Возможность использования языка JAVA
- Возможность анализа текста на основе регулярных выражений
- Повторное использование шаблонов

```
Phase: Phrase
Input: Token
Options: control = appelt


Rule: CheckThePhrase
(
 {Token.category  == JJ}
 {Token.category == NN}
):match
-->
:match.Phrase = {rule=CheckThePhrase}
```

Jape Transducer

Структура описания шаблона

```
Phase: Phrase
Input: Token
Options: control = appelt

Rule: CheckThePhrase
(
 {Token.category  == JJ}
 {Token.category == NN}
):match
```

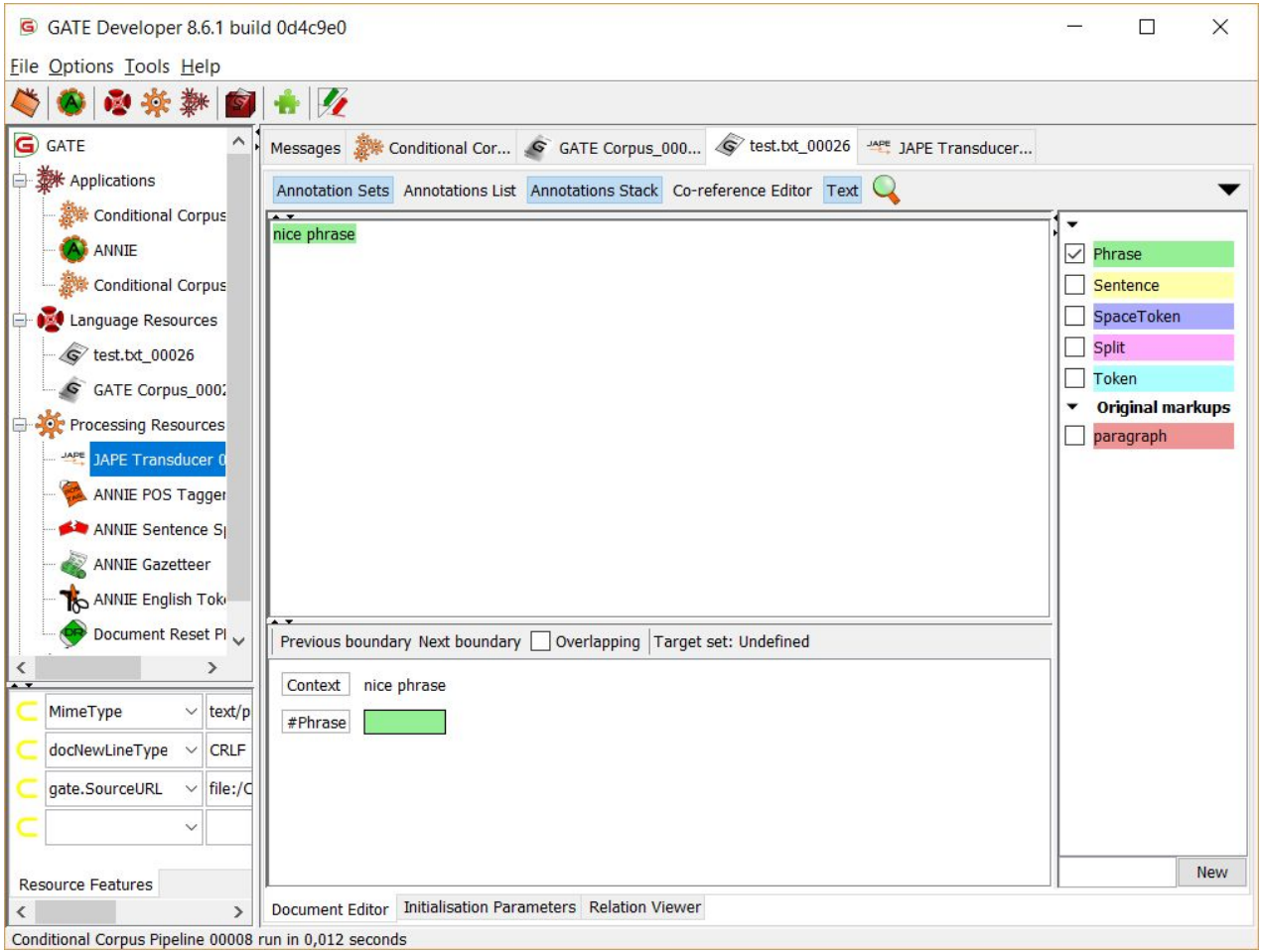
Идентификатор

-->

```
:match.Phrase = {rule=CheckThePhrase}
```

Действия для обнаружения шаблона

Операторы левой части	
|	Или
?	0 или 1 вхождение
*	0 или больше вхождений
+	1 или больше вхождений
=~	Регулярное выражение
Contains	Полностью включает
Within	Полностью находится в


GATE Developer 8.6.1 build 0d4c9e0
File Options Tools Help
GATE
Applications
Conditional Corpus
ANNIE
Conditional Corpus
Language Resources
test.txt_00026
GATE Corpus_000
Processing Resources
JAPE Transducer
ANNIE POS Tagger
ANNIE Sentence Sp
ANNIE Gazetteer
ANNIE English Tok
Document Reset Pl
MimeType
text/p
docNewLineType
CRLF
gate.SourceURL
file:/C
Resource Features
Messages
Conditional Cor...
GATE Corpus_000...
test.txt_00026
JAPE Transducer...
Annotation Sets
Annotations List
Annotations Stack
Co-reference Editor
Text
nice phrase
Phrase
Sentence
SpaceToken
Split
Token
Original markups
paragraph
Previous boundary
Next boundary
Overlapping
Target set: Undefined
Context
nice phrase
#Phrase
New
Document Editor
Initialisation Parameters
Relation Viewer
Conditional Corpus Pipeline 00008 run in 0,012 seconds

GATE

LSPL (Большакова Е.И.)

- Русскоязычные тексты
- учитываются морфологические характеристики и правила грамматического согласования
- Удобен в использовании лингвистами, не имеет привязки к языку программирования
- Повторное использование шаблонов

NP = A < c=nom, n=sing> N < c=nom, n=sing>

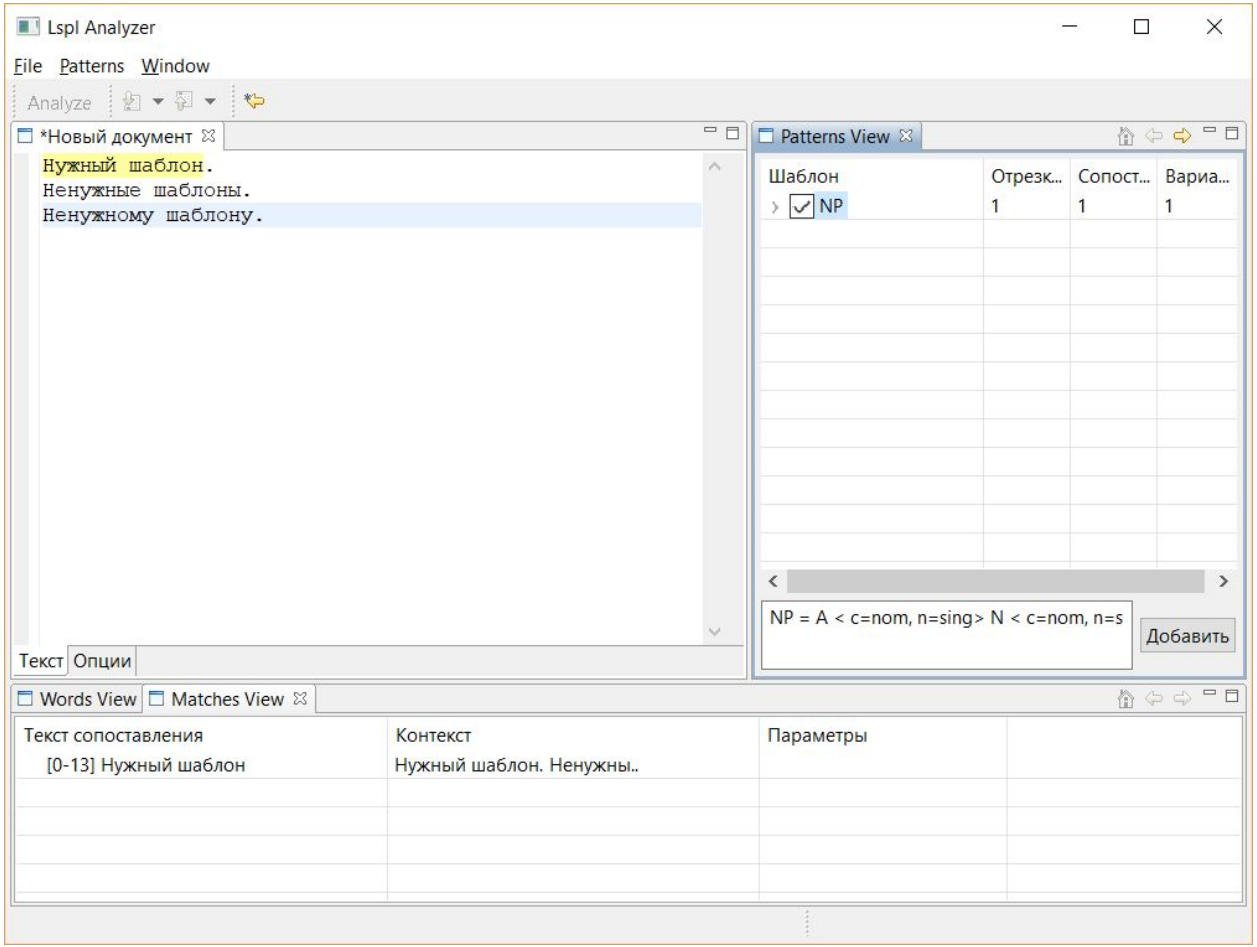
где A – прилагательное,
c=nom –именительный падеж,
n=sign – единственное число,
N – существительное.

Прилагательное + существительное единственного числа в именительном падеже

NP = A < c=nom, n=sing> N < c=nom, n=sing>

Операторы	
|	Или
[ ]	0 или 1 вхождение
<>	Условия (количество повторений, согласование)

Атрибуты	
c	Падеж
nom	Число
t	Время (у глаголов)



LSPL Analyser

LSPL (Рабчевский Е.А.)

- XML-подобный язык
- Тексты на английском и русском языке
- Шаблоны используются для автоматического построения онтологий
- Выходная семантическая модель - набор RDF триплетов (субъект, объект, предикат)

```
<pattern>
    <inputShema>
        <element type = “partOfSpeach” id=”1”>
            <content>adj</content>
        </element>
        <element type=”partOfSpeach” id=”2”>
            <content>noun</content>
        </element>
    </inputShema>
</pattern>
```

Атрибуты	
type	Тип элемента (literal, wordForm, partOfSpeech и syntacticGroup)
id	Идентификатор элемента в шаблоне по порядку
Элементы	
literal	Слово, указанное внутри : и, или, это
wordForm	Форма указанного слова
partOfSpeech	Слово указанной части речи
syntacticGroup	Синтаксическая группа, состоящая из нескольких слов, идущих подряд
punctualMark	Знак препинания. Если <content> пуст, подойдёт любой знак препинания

Томита

- Тексты на русском или английском языке
- Шаблон содержит две части – левую, состоящую из одного нетерминала и правую, состоящую из терминалов и нетерминалов
- Использование регулярных выражений
- Позволяет работать с грамматикой, словообразованием, словоформами и согласованием
- Повторное использование шаблонов
- Выходные форматы: xml, txt, protobuf, html

Операторы	
*	Символ повторяется 0 или более раз
+	Символ повторяется 1 или более раз
()	Символ входит в правило 0 или 1 раз
|	Используется для сокращенной записи правил с одинаковой левой частью
Пометы	
${ }	Значение подстановки
//	Комментарий
gram	Грамматические характеристики
gnc-agr	Согласование по роду
wfm, wff, wfl	Символ проверяется на соответствие регулярному выражению, указанному в поле помет

Терминалы	
Noun	Существительные (кроме имен, фамилий и отчеств)
Participle	Причастие
UnknownPOS	Нераспознанное морфологией слово
EOSent	Символ конца предложения
Граммемы	
ANUM	Порядковое числительное
APRO	Местоименное прилагательное
anim / inan	Одушевленность / неодушевленность
m / f / n / mf	Мужской / женский / средний / общий род
brev	Краткая форма прилагательных
1p / 2p / 3p	Первое / второе / третье лицо
Также можно определить степень сравнения, время, вид, залог, наклонение глагола и др.	

S -> Adj<nom, sg> Noun<nom, sg>,

где Adj - прилагательное,
Noun - существительное,
nom -именительный падеж,
sg - единственное число

```
Нужный шаблон.
Ненужные шаблоны.
Ненужному шаблону.
```

Нужный шаблон . **EOS**

Нужные шаблоны . **EOS**

Ненужному шаблону . **EOS**

Text	Type
Нужный шаблон	TAuxDicArticle [пример]

Сравнение

Язык	Предназначение	Работа с семантическими особенностями текста	Повторное использование шаблонов	Поддержка регулярных выражений	Система для работы с шаблонами
Jape	анализ текста на основе регулярных выражений	+	+	+	+ (Gate)
LSPL (Большакова Е.И.)	формальное описание конструкций русского языка с целью их представления в системах автоматической обработки русскоязычных текстов, основанных на морфологическом и частичном синтаксическом анализе	-	+	+	+ (LSPL Analyser)
LSPL (Рабчевский Е. А.)	семантический анализ текста, автоматическое построение онтологий	+	-	-	± (online анализатор и валидатор шаблонов)
Томита	извлечение структурированных данных (фактов) из текста на естественном языке	+	+	+	+ (Томита-парсер)

Библиографический список

1. GATE JAPE Grammar Tutorial Version 1.0 [Электронный ресурс] // URL: https://gate.ac.uk/sale/thakker-jape-tutorial/GATE%20JAPE%20manual.pdf.
2. Рабчевский Е.А. Автоматическое построение онтологий на основе лексико-синтаксических шаблонов для информационного поиска. // Труды 11-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL’2009. – Петрозаводск, 2009. – С. 69-77.
3. Большакова Е.И. Язык лексико-синтаксических шаблонов LSPL: опыт использования и пути развития // Программные системы и инструменты: Тематический сборник, № 15 / Под ред. Королева Л.Н. – М.: Изд. отдел факультета ВМК МГУ; МАКС Пресс, 2014. - С. 15-26.
4. Yandex/tomita-parser [Электронный ресурс] // URL: https://github.com/yandex/tomita-parser.