# Программная поддержка языка лексико-синтаксических шаблонов[1]

***Носков А.А.***
*Студент факультета вычислительной математики и кибернетики*
*Московский государственный университет имени М.В.Ломоносова, Москва, Россия*
*E–mail: alno87@mail.ru*

В последнее время в области создания систем автоматической обработки текстов на естественном языке активно развиваются средства [1], позволяющие специфицировать и находить в тексте различные языковые конструкции. К таким средствам относится язык LSPL [2], предложенный для описания конструкций русского языка (в первую очередь именных словосочетаний) и учитывающий его особенности, в частности, высокую флективность. Язык позволяет описывать языковые конструкции в виде лексико-синтаксических шаблонов, задающих словоформы, лексемы и их морфосинтаксические характеристики. Важной особенностью языка LSPL является возможность указывать грамматическое согласование между составными частями описываемых языковых конструкций. Например, шаблон *A N <A.g=N.g, A.n=N.n, A.c=N.c>* описывает сочетание из прилагательного (*A*) и следующего за ним существительного (*N*), согласованного с ним в роде (*g*), числе (*n*) и падеже (*c*). Язык LSPL также позволяет использовать ранее описанные шаблоны, что дает возможность сначала подготовить набор шаблонов, описывающих простые конструкции, а затем на их основе описывать более сложные конструкции.

Целью представленной в докладе работы было создание программной системы для поиска в тексте на русском языке конструкций, описанных в виде шаблонов на языке LSPL. При этом система должна была допускать простую интеграцию с другими программными средствами обработки текста, а также использоваться в виде отдельного приложения, поддерживающего работу пользователя-лингвиста по анализу текста.

В ходе реализации системы были выделены три основных компонента:

- ядро системы, осуществляющее частичный синтаксический анализ текста и поиск в нем конструкций, описанных в виде LSPL-шаблонов;
- программный интерфейс (API) для использования ядра из приложений на языке Java;
- графический пользовательский интерфейс для задания шаблонов, загрузки анализируемого текста и поиска в нем нужных конструкций.

Для более эффективной (с точки зрения производительности и использования памяти) реализации ядра использовался язык C++. Пользовательский интерфейс реализован на языке Java с использованием кроссплатформенной библиотеки SWT.

Ядро системы использует специальное графовое представление анализируемого текста, опирающееся на разметку его фрагментов с использованием аннотаций. Похожий подход для разметки текста был применен в системе GATE [1], однако предлагаемое графовое представление специализировано для эффективного поиска в нем конструкций на основе LSPL-шаблонов. Для оптимизации поиска используются заранее построенные индексы слов разных частей речи и индексы на основе возможных префиксов слов.

## Литература

1. Bontcheva K., et. al, Developing Reusable and Robust Language Processing Components for Information Systems using GATE // Proc. 13th Intern. Workshop on Database and Expert Systems Applications. IEEE Computer Society, Washington, DC, 2002, pp. 223-227
2. Большакова Е.И. и др. Лексико-синтаксические шаблоны в задачах автоматической обработки текстов // Труды межд. конф. Диалог '2007 – М.: Издательский центр РГГУ, 2007, с. 70-75.

[1] Работа выполнена при поддержке гранта РФФИ № 06-01-00571.