

# ЛЕКСИКО-СИНТАКСИЧЕСКИЕ ШАБЛОНЫ КАК СРЕДСТВО ИНЖЕНЕРНОЙ ЛИНГВИСТИКИ \*

Большакова Е.И.

Создание лингвистического обеспечения для систем автоматической обработки текстов (компьютерных словарей, тезаурусов, баз лингвистических данных) требует тщательного исследования лексических, синтаксических и дискурсивных особенностей текстов проблемной области с целью выявления состава и грамматических характеристик используемых в них языковых конструкций и последующей их формализации.

Существенную поддержку в проведении этих трудоемких задач инженерно-исследовательского характера оказывают программные средства, позволяющие автоматически находить в исследуемых текстах нужные единицы (словоформы, словосочетания и др.) и их контекст. К таким средствам можно отнести браузеры аннотированных корпусов текстов, осуществляющие обычно поиск слов и словоформ. Однако все чаще возникает потребность исследовать тексты, пока не включенные в корпусы, и при этом хотелось бы иметь возможность поиска более сложных единиц, чем словоформы, например, согласованных именных словосочетаний.

В настоящей работе обсуждается концепция лексико-синтаксического шаблона языковых конструкций и характеризуется основанный на ней декларативный язык лексико-синтаксических шаблонов (LexicoSyntactic Pattern Language), предложенный для описания лексических и грамматических свойств конструкций русского языка. Язык LSPL удобен:

- для запросов на поиск в текстах конструкций по их свойствам, сформулированным на основе их словарного состава и несложных грамматических условий;
- для формальной записи языковых конструкций с целью их представления в словарных компонентах систем автоматической обработки текстов.

При записи лексико-синтаксических шаблонов можно задавать конкретные словоформы и лексемы, а также произвольные слова нужной части речи, указывая при необходимости их грамматические характеристики (падеж, число, род, время, лицо и др.). В язык включены также конструкции повторения элементов шаблона и задания опциональных элементов, что делает его достаточно мощным. Важная особенность разработанного языка – возможность записи контекстно-зависимых условий, в первую очередь – грамматического согласования элементов шаблона, необходимого для описания именных словосочетаний русского языка.

Разработана библиотека программных методов сопоставления LSPL-шаблонов с текстом на основе его поверхностного синтаксического анализа. На базе библиотеки создан пользовательский интерфейс, позволяющий находить по заданному шаблону все соответствующие ему фрагменты исследуемого текста. Построен и исследован набор LSPL-шаблонов, отображающий характерные для научно-технических текстов конструкции (например, *Под T будем понимать D*) и включенный в состав системы автоматической обработки научно-технических текстов.

---

\* Работа выполняется при финансовой поддержке РФФИ (проект № 06-01-00571)