Формализация лексико-синтаксической информации для распознавания регулярных конструкций естественного языка

Большакова Е.И., Васильева Н.Э.

Решение многих прикладных задач автоматической обработки текста на естественном языке, таких как реферирование и аннотирование, извлечение знаний из текстов, литературно-научное редактирование, требует учета различных особенностей обрабатываемых текстов: лексико-фразеологических, синтаксических, логико-композиционных. Именно этот учет позволяет достичь приемлемого уровня решения указанных прикладных задач, базируясь на поверхностном синтаксическом анализе теста и избегая высокозатратного полного синтаксического разбора предложений [3, 4].

В целом, языковая специфика определяется функциональным стилем, жанром и конкретной предметной областью текста. Так, к характерным особенностям научно-технического стиля относится не только широкое использование специальных терминов, но и особый способ изложения – рассуждение, призванное объяснить и обосновать полученные результаты. К типичным шагам рассуждения относится введение термина, приведение фактов и доказательств, подведение итогов и др. Эти шаги организуются в тексте при помощи общенаучных слов и выражений (*определим как*, *в заключение*, *по причине того, что* и т.д.), из которых формируются фразы-клише научной прозы: *будем далее считать*, *из вышесказанного следует*, *как показал проведенный опыт* и т.п. [5].

Для автоматического распознавания в текстах подобных языковых выражений не достаточно обычной словарной информации – необходимо взаимосвязанное представление характеризующей их лексической и синтаксической информации. Нами было предложено записывать эту информацию в виде специальных декларативных структур, названных ***лексико-синтаксическими шаблонами*** [1]. По сути, лексико-синтаксический шаблон – это структурный образец языковой конструкции, который отображает ее лексические и поверхностно-синтаксические свойства. Наша концепция лексико-синтаксического шаблона идейно близка к работам [6, 7], однако развита с учетом специфики русского языка и воплощена в специально разработанном языке шаблонов LSPL. На базе этого языка была проведена формализация характерных для научно-технических текстов конструкций определений новых терминов.

В данной работе обсуждается назначение декларативного языка лексико-синтаксических шаблонов, характеризуются его выразительные возможности, и показывается его применимость для отображения регулярных конструкций русского языка.

Задачи формализации характерных языковых конструкций

В ходе разработки процедур автоматической обработки русских научно-технических текстов нами были изучены синтаксические и лексико-фразеологические особенности текстов этого стиля, взятых из разных предметных областей (строительство, информатика и вычислительная техника и др.) [1, 2, 5]. Это позволило обнаружить множество типичных для научной прозы выражений (*Под T будем понимать D*, *Предположим, что S* и т.д.) и показало их важность для организации связного текста. Такие конструкции следует формализовать и представить в словаре системы автоматической обработки научно-технического текста для реализации более полного и глубокого терминологического анализа, а также распознавания логико-композиционной структуры текста.

Было выявлено, что регулярные языковые конструкции содержат фиксированные лексемы и имеют определенную синтаксическую структуру. Для автоматического распознавания такие выражения удобно описать в виде специальных лексико-синтаксических шаблонов. Каждый лексико-синтаксический шаблон содержит конкретные словоформы (*понимать*, *что*, *предположим* и т.п.) и свободные места (слоты), заполняемые определенными синтаксическими конструкциями. К примеру, шаблон `под Tins будем понимать NGacc` содержит совместно встречающиеся слова *будем*

и *понимать*; элемент *Tins* обозначает новый вводимый термин, который должен быть выражен согласованной (в роде, числе и падеже) именной группой с главным словом в форме творительного падежа; элемент *NGacc* заполняется согласованной именной группой в винительном падеже, которая выражает определение или пояснение вводимого термина. Указанный шаблон описывает, в частности, такую фразу: *Под семантической связью будем понимать отношение понятий в понятийной системе предметной области*.

Формализация каждой регулярной конструкции в виде шаблона предполагает определение множества входящих в нее лексем и их возможных грамматических форм, а также выявление необходимых синтаксических условий. Эта работа была в первую очередь проделана для конструкций определений новых (авторских) терминов [2]. Вручную было просмотрено около 70 научно-технических текстов, и из них были выделены фразы, которые использовались при определении нового термина. После их предварительного анализа было получено первоначальное множество конкретных лексем, входящих в конструкции определений, что позволило в дальнейшем с помощью обычного текстового редактора выявить новые фразы определений терминов.

Так как разнообразие найденных фраз определений терминов было довольно велико, они были сгруппированы по используемым в них одной-двум общим лексемам, и далее полученные группы фраз рассматривались по отдельности, что позволило выявить их грамматические особенности и формализовать группы в виде набора лексико-синтаксических шаблонов. Этот набор покрывает в совокупности примерно 60-70% определений терминов, встречающихся в русских научно-технических текстах.

Ясно, что проведение подобной работы для выявления и формализации даже базового множества характерных для научной прозы конструкций – достаточно трудоемкая задача и желательно автоматизировать ее решение. Существенную помощь могло бы оказать специальное программное средство, позволяющее автоматически находить в текстах фрагменты с исследуемыми конструкциями по частичному описанию их лексико-синтаксических свойств. Такое описание можно было бы задавать в виде лексико-синтаксического шаблона, используя его как отправную точку для дальнейшего уточнения возможных вариантов, состава и синтаксических особенностей формализуемых конструкций.

Тем самым, одной из задач наших исследований стала разработка формального языка записи шаблонов, который мог бы использоваться как:

– способ формальной записи специфических языковых конструкций для их представления в системе автоматической обработки научно-технических текстов;

– язык записи запросов на поиск исследуемых конструкций в текстах, формулируемых на основе входящих в них слов и несложных грамматических условий.

При решении этой задачи мы старались отобрать в создаваемый язык выразительные средства, позволяющие достаточно гибко записывать лексикографические единицы (символьные строки, словоформы, лексемы) и их морфологические характеристики, причем в как можно более простой и явной форме. Мы исходили из того, что включаемые в язык средства записи грамматических характеристик должны быть понятны не только лингвистам, но и другим специалистам, участвующим в разработке шаблонов. Другое существенное требование к языку – возможность явного задания связи синтаксического согласования (типичной для русских именных словосочетаний), которая отсутствует в известных формальных языках описания специфических языковых конструкций (например, [3]) .

Основные возможности языка LSPL

Лексико-синтаксический шаблон состоит из имени и тела, разделяемых знаком равенства. В общем случае тело шаблона определяет последовательность элементов, из которых должна состоять описываемая языковая конструкция, и задает условия грамматического согласования этих элементов. К примеру, шаблон *AN = A N <A=N>*

имеет имя *AN* и тело из элементов *A*, *N* и условия согласования *A=N*. Этот шаблон описывает именную группу из прилагательного (*A*) и существительного (*N*), согласованную по всем их морфологическим параметрам (падеж, число, род).

Основными элементами шаблона являются элемент-строка и элемент-слово. ***Элемент-строка*** позволяет описать в виде символьной строки (в двойных кавычках) конкретную словоформу, сокращение или знак препинания: *"рамой"*, *"т.е."*, *"–"*.

Элемент-слово соответствует отдельному слову описываемой языковой конструкции, для которого в общем случае указываются:

- часть речи (используются известные символьные обозначения: N – существительное, V – глагол, A – прилагательное, Pr – предлог, Pn – местоимение и т.д.);
- конкретная лексема, определяющая множество всех словоформ слова;
- значения морфологических параметров слова, сужающие множество допустимых словоформ (параметры записываются в угловых скобках после лексемы; их обозначения: c – падеж, n – число, g – род, t – время, p – лицо и т.д.).

К примеру, элемент-слово *V<пониматься; t=pres, p=3, m=ind>* описывает глагол *пониматься* в формах настоящего времени 3 лица изъявительного наклонения, т.е. задает его словоформы *понимается* или *понимаются*. При задании элемента-слова конкретная лексема и значения морфологических параметров могут быть не указаны, что позволяет задать любую словоформу данной лексемы (например: *N<файл>*) или же произвольное слово определенной части речи с нужными грамматическими характеристиками (например, *A<; c=ins, n=sing>* задает любое прилагательное в форме творительного падежа единственного числа).

В общем случае в шаблон могут входить как несколько элементов-слов разных частей речи, так и несколько разных слов одной части речи; для их различения можно использовать числовые индексы, например, шаблон *NN = N1 N2<; c=gen>* включает два различных существительных *N1* и *N2*, причем второе из них – в родительном падеже.

Условия согласования описывают связь синтаксического согласования отдельных элементов шаблона и относятся ко всему шаблону в целом. Они задаются после описания всех элементов шаблона, в виде равенства значений согласуемых морфологических признаков (в угловых скобках, подобно конкретизации морфологических параметров). К примеру, шаблон *PnV = Pn V <Pn.n=V.n, Pn.g=V.g>* описывает согласованные (в числе и роде) пары из местоимения и глагола: *мы введем*, *они разработали*, *я ищу* и т.д.

В шаблоне может быть задано ***повторение элементов***, которое записывается с помощью фигурных скобок: в них указываются элементы, которые могут встречаться в тексте несколько раз подряд. Например, повторение *{N<; c=gen>}* задает цепочку из идущих подряд существительных в родительном падеже. Если известны ограничения на количество одинаковых элементов, то их можно указать в шаблоне: запись *{A}<1,3> N* задает последовательность из одного, двух или трех прилагательных и существительного.

Язык LSPL позволяет включать в шаблон ***опциональные элементы*** (в квадратных скобках): например, элемент *["не"]* указывает необязательность вхождения частицы *не* в описываемое языковое выражение. Допустима и запись ***альтернативных вариантов*** некоторой языковой конструкции, для чего используется символ |. К примеру, шаблон *AP = A|Pa* описывает понятие адъектива, т.е. прилагательного (*A*) или причастия (*Pa*).

Лексико-синтаксический шаблон может включать ***параметры***, которые записываются в круглых скобках после всех его элементов и фиксируют те или иные неконкретизированные (т.е. не имеющие значения) морфологические параметры его элементов. Например, параметрами шаблона *AAN = A1 A2 N <A1=A2=N> (N)* являются все морфологические характеристики элемента-слова *N* (шаблон задает именную группу, согласованную по всем общим для ее элементов морфологическим параметрам).

В качестве элемента шаблона может быть использован другой, ранее описанный шаблон, т.е. ***экземпляр шаблона***. Он задается именем используемого шаблона и

последующими конкретизациями его характеристик-параметров (в угловых скобках). К примеру, шаблон *TD17 = "далее" "–" NG<;c=nom>* описывает языковые фразы, в которых после слова *"далее"* через тире идет именная группа (*NG*) в именительном падеже (например: *далее – базовый алгоритм*). В этом шаблоне как экземпляр использован шаблон с именем *NG*, для которого конкретизирован падеж – именительный (*c=nom*).

В свою очередь шаблон *NG = {A1} N1 {N<;c=gen>} <A1=N1> (N1)* состоит из существительного *N1* (главного слова), последовательности согласованных с ним прилагательных *{A1}* и цепочки существительных в родительном падеже *{N<;c=gen>}* (например: *восходящий процесс порождения элементов решетки*). Параметр шаблона *N1* означает, что группа *NG* наследует весь набор морфологических характеристик главного существительного *N1* – это позволяет использовать параметр при конкретизации морфологических характеристик группы *NG* в шаблоне *TD17*.

Таким образом, при создании шаблона сложной языковой конструкции имеет смысл выделить ее составные части и описывать их по очереди в виде шаблонов.

Использование LSPL-шаблонов языковых конструкций

Разработанный язык был применен в первую очередь для создания шаблонов регулярно используемых в научно-технических текстах фраз-определений новых терминов. В таблице 1 приведены примеры полученных шаблонов, иллюстрирующие декларативный характер языка и его выразительные возможности (последний пример получен при формализации типичных фраз методических документов деловой прозы). Во всех примерах участвует экземпляр вышеописанного шаблона *NG*, который представляет собой один из наиболее распространенных синтаксических образцов терминов научно-технической и деловой прозы. В шаблоне *TD18* (четвертый пример) использован экземпляр шаблона с именем *Ab*, задающего акроним (т.е. инициальную аббревиатуру).

Таблица 1. Шаблоны регулярных языковых конструкций

Шаблон	Пример фразы
TD2 = NG1<; c=ins> V<называться; t=pres, p=3, m=ind> NG2<; c=nom> [PaG] <NG1.n=V.n=NG2.n, PaG=NG2>	*Трансформационным признаком называется приоритетный признак, выделяющий некоторые именные группы в предложении*
TD6 = NG1<; c=acc> ["мы"] "будем" "называть" NG2<; c=ins> <NG1.n=NG2.n>	*Поэтому эту операцию будем называть правилом генерализации примеров*
TD25 = "под" NG1<; c=ins> V<пониматься; t=pres,p=3, m=ind> NG2<;c=nom> <NG1.n=NG2.n>	*…под синтаксемой понимается такое дерево, в корне которого стоит существительное …*
TD18 = NG "(далее"["–"]Ab<;c=nom> ")"	*… все концепты области-источника (далее ОИ), …*
AD1 = NG1<;c=nom> Pa<разработанный; f=short> "в" "целях" NG2<; c=gen> <NG1.n=Pa.n, NG1.g=Pa.g>	*Методика планирования себестоимости услуг разработана в целях обеспечения единства состава и классификации затрат…*

На основе лексико-синтаксических шаблонов может выполняться распознавание в тексте регулярных конструкций и выделение их значимых частей. Рассмотрим пример обработки фразы *Трансформационным признаком называется приоритетный признак, выделяющий некоторые именные группы в предложении* (см. Рисунок 1). Во фразе встречается слово *называется*, входящее в состав нескольких шаблонов, но поскольку перед ним расположена согласованная именная группа с главным словом в творительном падеже, а после него следует именная группа в именительном падеже, будет выбран шаблон, представленный в первой строке Таблицы 1. После успешного наложения шаблона (т.е. проверки записанных в нем синтаксических условий) из фразы будет извлечен термин *трансформационный признак* и его определяющее выражение.

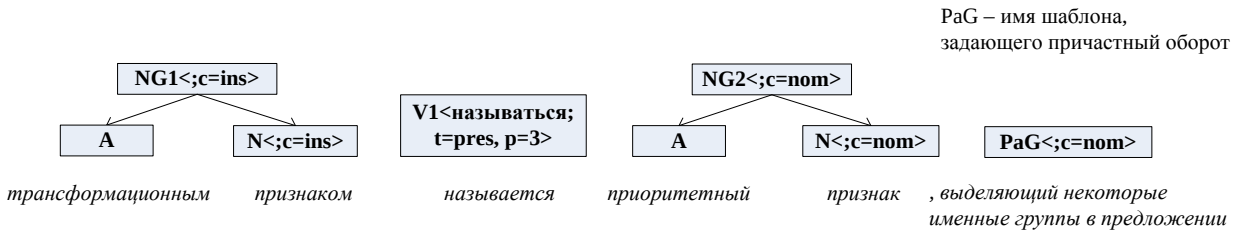

Рисунок 1. Схема наложения шаблона

Язык LSPL был применен при разработке словарных компонентов системы автоматической обработки научно-технических текстов, для формального описания:

– регулярных языковых конструкций определений новых (авторских) терминов;

– синтаксических образцов научно-технических терминов и их синонимичных вариантов (например, *библиотека стандартных программ – библиотека программ*);

– случаев объединений в тексте нескольких терминов (к примеру, фрагмент *ЭВМ второго, третьего и четвёртого поколения* представляет объединение терминов *ЭВМ второго поколения*, *ЭВМ третьего поколения* и *ЭВМ четвёртого поколения*).

В целом, язык LSPL пригоден для задания любой лексической и поверхностно-синтаксической информации, на основе которой можно распознавать регулярные языковые конструкции. Представление такой информации в системах автоматической обработки текстов позволит осуществлять более широкий спектр интеллектуальных операций над текстом. В настоящий момент завершается разработка библиотеки программных компонентов, поддерживающих распознавание в тексте на естественном языке конструкций по заданным LSPL-шаблонам.

Работа выполнена при финансовой поддержке РФФИ (проект № 06-01-00571).

Литература

1. Большакова Е.И., Васильева Н.Э., Морозов С.С. Лексико-синтаксические шаблоны для автоматического анализа научно-технических текстов // Десятая Национальная конференция по искусственному интеллекту с международным участием КИИ-2006. Труды конференции в 3-х томах. Т. 2. – М.: Физматлит, 2006, с.506-524.
2. Васильева Н.Э. Шаблоны употреблений терминов и их использование при автоматической обработке научно-технических текстов // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог ‘2004 – М.: Наука, 2004, с. 96-101.
3. Ермаков А. Е., Плешко В.В., Митюнин В.А. RCO Pattern Extractor: компонент выделения особых объектов в тексте // Информатизация и информационная безопасность правоохранительных органов: XII Международная научная конференция. Сборник трудов – Москва, 2003, с. 312-317.
4. Хорошевский В.Ф. OntosMiner: семейство систем извлечения информации из мультиязычных коллекций документов // Девятая Национальная конференция по искусственному интеллекту с международным участием КИИ-2004. Т. 2. – М.: Физматлит, 2004, с.573-581.
5. Bolshakova, E.I. Lexicon of Common Scientific Words and Expressions for Automatic Discourse Analysis of Scientific and Technical Texts // Proceedings of the XIII-th Int. Conference “Knowledge-Dialogue-Solution”, V. 2. Sofia: FOI ITHEA, 2007, p. 551-558.
6. Hearst, M.A. Automated Discovery of WordNet Relations. In: Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database. MIT Press, Cambridge, 1998, p.131-151.
7. Paice, C., Jones P. (1993) The Identification of Important Concepts in Highly Structured Technical Papers. Proc. of 16th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburg, 1993, p.69-78.