Шаблоны употреблений терминов и их использование при автоматической обработке научно-технических текстов

Н. Э. Васильева
МГУ им. М.В.Ломоносова, Факультет ВМиК
nvasil@port.ru

В работе описываются результаты исследования лексических контекстов употреблений терминов в научно-технических текстах, проведенного с целью учета лексико-стилистических особенностей употребления терминов в текстах и их дальнейшего использования для повышения надежности и эффективности автоматического выделения терминов. В первую очередь анализировались конструкции, вводящие новые (авторские) термины, рассматривались также контексты употреблений общепринятых (словарных) терминов. В докладе приводятся собранные статистические данные о контекстах, встреченных в проанализированных текстах. Рассматриваются примеры лексико-синтаксических шаблонов, полученных в результате формализации наиболее типичных контекстов. Кратко характеризуется формальный язык, предложенный для записи указанных шаблонов. Также обсуждаются способы использования лексико-синтаксических шаблонов при автоматическом анализе терминологического состава научно-технического текста.

Введение

Задача анализа терминологического состава текста является ключевой при решении многих прикладных задач автоматической обработки текстов, таких как индексирование и рубрикация документов, создание терминологических словарей и тезаурусов. Несмотря на интенсивное изучение данной задачи, она решена далеко не полностью, и средства автоматического выделения терминов в текстах до сих пор не включены в состав общедоступных лингвистических технологий наравне, например, с морфологическими анализаторами.

Во многих приложениях, например, при автоматическом индексировании документов или автоматизированном литературно-научном редактировании текстов, необходимо проведение как можно более полного и глубокого терминологического анализа текста, что предполагает распознавание в тексте всех терминоупотреблений, а именно: появлений новых терминов и их дальнейших употреблений [4, 5], употреблений в тексте общепринятых терминов [3] и т.д. Для повышения эффективности и надежности такого распознавания требуется привлечение не только обычной словарной информации [1], но и информации о лексических и синтаксических особенностях употребления терминов в

текстах определенного стиля, например, для научно-технической прозы характерно использование регулярных конструкций, определяющих термины.
В докладе описываются результаты исследования лексических контекстов употреблений терминов в научно-технических текстах. Исследования проводились на коллекции научно-технических текстов (около 250 текстов), включающей, в том числе, тексты докладов международного семинара Диалог 2002. В ходе исследования рассматривались характерные для научно-технической прозы конструкции, используемые при определении новых терминов [5], а также другие контексты терминоупотреблений. В результате было собрано и проанализировано около 40 контекстов употреблений терминов.
В работе приведена соответствующая статистическая информация о встречаемости различных контекстов употреблений терминов в научно-технических текстах и рассматриваются примеры использования наиболее типичных контекстов.
Для применения выявленной контекстной информации при автоматическом распознавании терминов была проведена формализация контекстов и предложен язык описания контекстов употреблений терминов, фрагменты которого кратко характеризуются в докладе. Также обсуждаются возможные способы использования лексико-синтаксических шаблонов, полученных при формализации контекстов употреблений терминов.

1. Контексты употреблений терминов

Под терминами обычно понимают устоявшиеся в своем значении слова и словосочетания, обозначающие понятия некоторой проблемной области. Такие термины принадлежат уже сложившейся терминологической системе и зафиксированы в соответствующем терминологическом словаре; поэтому далее будем называть их словарными [4, 5]. В научно-технических текстах довольно часто встречаются термины, не являющиеся общепринятыми и отсутствующие в терминологических словарях. В большинстве случаев они определяются автором текста и выражают новые понятия, формулируемые и описываемые в этом тексте. Такие термины будем называть авторскими [4, 5].
Авторские термины нередко встречаются в контексте их определения или пояснения, например, фраза *«Здесь селективностью мы называем то, что семантика слов, которые…»* определяет термин *селективность*, а фраза *«Под глагольной фразеологической единицей (далее ГФЕ) мы понимаем такую фразеологическую единицу, в которой...»* вводит термин *глагольная фразеологическая единица* и синоним для него *ГФЕ*.
Кроме контекстов определений новых терминов для научно-технической прозы характерны и другие контексты употреблений терминов – как словарных, так и авторских. В частности, контексты могут раскрывать семантические связи между терминами [3]. Например, фраза *«…совокупность разнотипных данных, используемых в одном контексте, будем называть динамическим срезом информационного пространства или сценарием»* вводит новый термин (*динамический срез информационного пространства*) и выражает связь синонимии между терминами *динамический срез информационного пространства* и *сценарий*; а фраза *«…наиболее формальными графическими образами являются графики аналитических зависимостей»* выражает связь «род-вид» между терминами *графический образ* и *график аналитических зависимостей*.
Как показало проведенное нами исследование, термины обычно употребляются в составе конструкций, содержащих определенные лексемы и имеющих определенную синтаксическую структуру. Например, фраза *«Под графемной конструкцией понимается графическая форма...»* схематически может быть описана как

«под» T_{INS} «понимается» NG_{NOM},

где кавычками выделяются фиксированные словоформы; T_{INS} − авторский термин, выраженный согласованной именной группой, главное слово которой имеет форму творительного падежа; NG_{NOM} − описание или объяснение авторского термина, выраженное согласованной именной группой, главное слово которой имеет форму именительного падежа.

2. Исследование контекстов употреблений терминов

Исследование было начато с изучения контекстов определений авторских терминов. Для этого вручную было просмотрено около 50 научно-технических текстов, и из них были выделены те конструкции, которые использовались при определении или пояснении нового термина. После их предварительного анализа было получено первоначальное множество определенных лексем, входящих в конструкции определений, что позволило в дальнейшем частично автоматизировать процесс поиска новых контекстов определений терминов. Выделение контекстов, содержащих уже найденные лексемы, осуществлялось с помощью операций поиска обычного текстового редактора.

Так как количество разных контекстов быстро увеличивалось, было принято решение разделить контексты на группы так, чтобы каждой определенной лексеме (или двум-трем совместно встречающимся лексемам) соответствовала своя группа контекстов. Например, к одной группе были отнесены конструкции вида

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

где «мы» и «будем называть» − совместно встречающиеся лексемы, причем слово «мы» может отсутствовать; T_{INS} − авторский термин, выраженный согласованной именной группой, главное слово которой имеет форму творительного падежа; NG_{ACC} − описание или объяснение авторского термина, согласованной именной группой, главное слово которой имеет форму винительного падежа, возможно расширенной придаточным предложением. Такой контекст описывает, например, фразу *«…значение, которое используется для расширения первоначального набора - не теста, мы будем называть существенным значением…»* и фразу *«Такие операции будем называть понятийными операциями…»*.

После разделения контекстов на группы был проведен анализ контекстов каждой группы, который показал, что одни и те же определенные лексемы используются в составе нескольких синтаксически разных конструкций. Например, слово *называться* используется как минимум в 2 различных контекстах определений новых терминов (см. Таблицу 1).

На данный момент выделено около 20 групп. В Таблице 1 рассмотрены примеры групп и входящих в них контекстов. В квадратные скобки заключены факультативные элементы; авторские термины (термины-кандидаты) подчеркнуты.

Таблица 1. Примеры групп контекстов

Контексты группы	**Синтаксические условия**	**Примеры**
Группа *называться*		
T_{INS} «называться» NG_{NOM}	глагол «называться» в форме 3 лица настоящего времени; число термина-кандидата T_{INS} и глагола совпадают	*Трансформационным признаком называется приоритетный признак…*
NG_{NOM} «называться»	глагол «называться» в форме	*… порождающие грамматики*

T_{INS}	3 лица настоящего времени; число T_{INS} и глагола совпадают	*называются репродукционными грамматиками*
Группа *называть*		
T_{INS} P «называть» NG_{ACC}	глагол «называть» в форме насто-ящего времени; лицо и число мес-тоимения P и глагола совпадают; число T_{INS} и главного слова именной группы NG_{ACC} совпадает	*... «инкорпорацией участника» мы называем такое семантическое соотношение двух лексических единиц…*
NG «, который» P «называть» T_{INS}	глагол «называть» в форме насто-ящего времени; лицо и число мес-тоимения P и глагола совпадают; род, число, падеж местоименного прилагательного «который» и NG совпадают	*...распространен в информационных системах метод, который я называю методом «встречного текста»*
Группа *так + называемый*		
«так называемый» T	род, число, падеж T и причастия «называемый» совпадают	*... предлагается ввести - так называемую иерархическую весовую функцию (ИВФ)...*
Группа *понимать*		
«под» T_{INS} [«мы»] «понимать» NG_{ACC}	глагол «понимать» в форме первого лица множественного числа настоящего времени	*...в данной статье под ЕЯ-интерфейсом понимается не диалоговая система, а система...*
Группа *быть + понимать*		
«под» T_{INS} [«здесь»] «быть» «понимать» NG_{ACC}	глагол «быть» в форме первого лица множественного числа будущего времени	*Под объектом здесь будем понимать любые исходные данные...*
Группа *далее*		
«(далее» [«–»] T_{NOM} «)»		*(далее – показатель «прямые и близкие попадания»)*

В ходе анализа собранной информации выяснилось, что часть выделенных контекстов не являются контекстами определений: некоторые контексты задают, например, отношение синонимии между понятиями. К примеру, предложение *«Гамма-знак будем также называть свободным стимулом (free stimulus)»* вводит синоним *свободный стимул* для авторского термина *гамма-знак*.

В настоящее время выделены лексемы (около 15), которые входят в контексты, не являющиеся контекстами определений (например, *являться*, *входить*, *часть*, *состав*, *использоваться*). Дополнительно были изучены материалы, содержащие информацию о способах выражения семантических связей между терминами и сочетаемости терминов с другими словами в предложении [2, 3]. Планируется использовать полученную информацию при дальнейшем исследовании контекстов употреблений терминов.

3. Формализация контекстов

Для использования собранных контекстов определений в системах автоматической обработки текстов они были формализованы, т.е. описаны в виде специальных лексико-синтаксических шаблонов, подобных тем, что предложены в [6]. В состав шаблонов входят следующие элементы:

- Литералы, т.е. конкретная словоформа, сокращение или знак препинания (например, «т.н.», «–», «под»). Литеральные элементы заключаются в кавычки.
- Символьные обозначения слов определенной части речи и синтаксических конструкций. Например, V– глагол, Pa – причастие, T – термин-кандидат, выраженный простой или сложной именной группой (сложная именная группа может содержать союзы, группы с предлогами, скобочные конструкции, придаточные предложения, причастные обороты), TA – часть термина-кандидата, выраженная одним или несколькими прилагательными, объединенными с помощью союза, TN – часть термина-кандидата, выраженная простой или сложной именной группой, Ng – простая или сложная именная группа.
- Условия, уточняющие грамматические характеристики слов и конструкций (например, Ng.number=V.number – число Ng и V совпадают, person= «3» – третье лицо).

Далее рассмотрим примеры шаблонов, полученные при формализации контекстов.

Пример 1.

Шаблон **«под» T<:case=«ins»> V<«пониматься»:tense=«pres» person=«3»> Ng<:case=«nom»> <T.number=V.number>**

описывает случаи вида *«Под графемной конструкцией понимается графическая форма, построенная из базисных, проблемно-ориентированных и/или графических конструкций»* и *«Под данными при такой формализации понимаются последовательности символов (слова, предложения) в некоторых алфавитах»* (см. Рисунок 1).

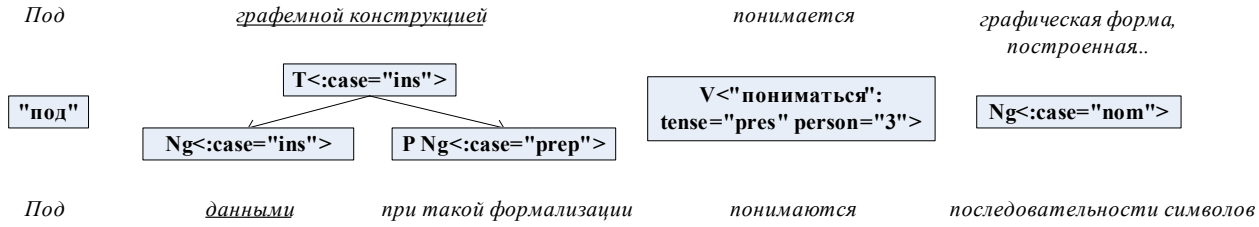

Рисунок 1. Схема применения шаблона.

Пример 2.

Шаблон **Ng «,» Pa<«названный»> T<:case=«ins»> <Ng.case=Pa.case Ng.number=Pa.number=T.number Ng.gender=Pa.gender>**

описывает случаи вида *«По результатам генерации форм, слова были разбиты на группы, названные профилями»*. В то же время, фраза *«...устойчивого выражения, названного в заголовке, в левой (объясняемой) части словарной статьи»* не вводит новый термин и не удовлетворяет шаблону, т.к. после причастия *«названный»* не следует конструкция, имеющая характеристики термина-кандидата.

В конце настоящей статьи приведена итоговая Таблица 2, содержащая информацию о формализованных контекстах, частоте их встречаемости в проанализированной коллекции текстов (процентное соотношение количества текстов, в которых встретился контекст, к количеству проанализированных текстов) и степени успешности (процентное соотношение случаев действительного введения авторских терминов к общему количеству фраз, в которых встретился контекст).

4. Использование лексико-синтаксических шаблонов

На наш взгляд предложенные лексико-синтаксические шаблоны могут быть использованы для:

- выявления в тексте терминов-кандидатов (авторских терминов) и определяющих или поясняющих их языковых выражений;
- распознавания синонимов для авторских терминов;
- определения предполагаемых семантических связей между авторскими и словарными терминами.

Шаблоны позволяют учитывать случаи сложного вхождения авторских терминов: объединение терминов с помощью сочинительных союзов и знаков препинания, а также разрывы терминов. Рассмотрим пример выделения терминов из фразы «Кванторы общности, не являющиеся ни кванторами основания, ни долевыми кванторами, называются немаркированными или недистрибутивными» (см. Рисунок 2).
В данной фразе встретилось слово «называются», входящее в состав нескольких шаблонов (см. Таблицу 1). Т.к. до глагола «называются» расположена согласованная именная группа, расширенная причастным оборотом, главное слово которой имеет форму именительного падежа, а после глагола следуют прилагательные, объединенные союзом «или», имеющие форму творительного падежа, будет выбран шаблон
TN<:case= «nom»> V<«называться»:tense=«pres» person=«3»> TA<:case=«ins»> <TN.number=V.number=TA.number TN.gender=TA.gender>.
После наложения шаблона (проверки синтаксических условий) и извлечения соответствующих элементам шаблона языковых конструкций мы получим два синонимичных термина-кандидата: немаркированный квантор общности и недистрибутивный квантор общности.

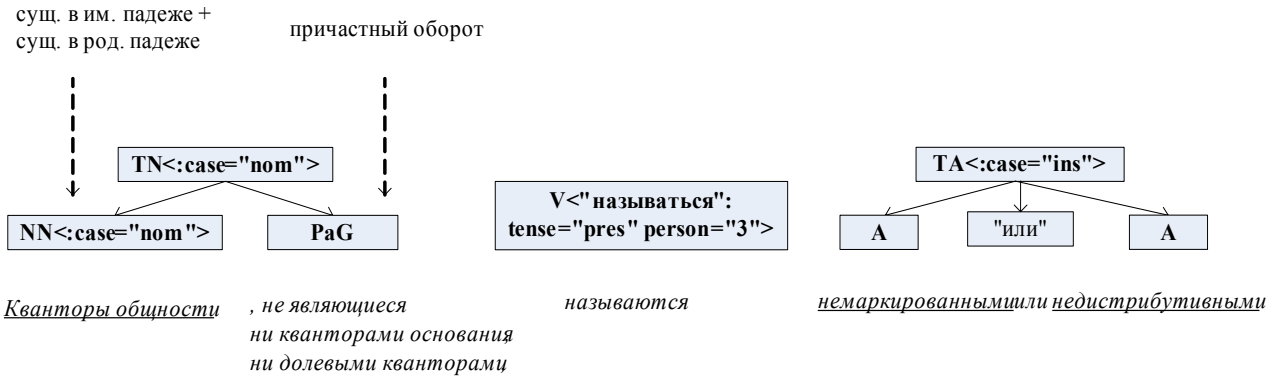

Рисунок 2. Схема применения шаблона.

В ближайшем будущем планируется программно реализовать основанный на шаблонах алгоритм выделения в тексте авторских терминов.

Литература

1. Большакова Е. И. О принципах построения компьютерного словаря общенаучной лексики //Труды Международного семинара Диалог ‘2002 по компьютерной лингвистике и интеллектуальным технологиям. М., 2002, Т. 1, с. 19-23.
2. Ершов А.П., Шанский Н.М., Окунева А.П., Баско Н.В. Терминологический словарь по основам информатики и вычислительной техники. - М., Просвещение, 1991.
3. Никитина С. Е. Тезаурус по теоретической и прикладной лингвистике. – М., Наука, 1978.
4. Пшеничная Л.Э., Коренга О.Н. Научный термин в словаре и тексте //НТИ. Сер.2. 1991, №12, с. 2-7.

5. Bolshakova, E. Recognition of Author's Scientific and Technical Terms. In: Computational Linguistics and Intelligent Text Processing. Second Int. Conf. CICLing 2001. A. Gelbukh (Ed.). Lecture Notes in Computer Science, N 2004, Springer-Verlag, 2001, p. 281-290.

6. 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.

Таблица 2. Примеры шаблонов определений

Шаблон	**Пример**	**% встреч.**	**% успеш.**
Группа <u>называться</u>			
TN<:case=«nom»> V<«называться»:tense=«pres» person=«3»> TA<:case=«ins»> <TN.gender=TA.gender TN.number=V.number=TA.number>	*<u>Вариант Т</u> называется <u>предпочтительным</u>, если он является ...*	6,7	50
T<:case=«ins»> V<«называться»:tense=«pres» person=«3»> Ng<:case=«nom»> <T.number=V.number=Ng.number>	*<u>Трансформационным признаком</u> называется приоритетный признак...*	3,1	25
Группа <u>называть</u>			
T<:case=«ins»> P V<«называть»:tense=«pres»> Ng<:case=«acc»><P.person=V.person P.number=V.number Ng.number=T.number>	*Здесь <u>селективностью</u> мы называем то, что семантика слов...*	3,3	100
Группа <u>быть+ называть</u>			
Ng<:case=«acc»> [«мы»>] V<«быть»:tense=«fut» person=«1» number=«plur»> «называть» T<:case=«ins»> <Ng.number=T.number>	*Такие операции будем называть <u>понятийными операциями...</u>*	5	100
Группа <u>мочь + быть + назван</u>			
V<«мочь»:tense=«pres» person=«3» number=«sing»> «быть» Pas<«назван»> T<:case=«ins»> <Pas.gender=T.gender Pas.number=T.number>	*Этот принцип обобщения ... может быть назван <u>гомоморфизмом понятий</u>*	6,8	100
Группа <u>назвать</u>			
Ng<:case=«acc»> [«мы»] V<«назвать»: tense=«pres» person=«1» number=«plur»> T<:case=«ins»> <Ng.number=T.number>	*Эту операцию назовём <u>операцией поиска существенных примеров</u>*	7,2	80
Группа <u>можно + назвать</u>			
можно» [«было» «бы»] [P<:person=«3»>] «назвать» T<:case=«ins»> <[T.number=P.number T.gender=P.gender]>	*Можно было бы назвать их <u>контекстно-нечеткими классами</u>*	10	71,4
Группа <u>так + называемый</u>			
«так» Pa<«называемый»> T <Pa.case=T.case Pa.gender=T.gender Pa.number=T.number:T>	*...предлагается ввести - так называемую <u>иерархическую весовую функцию (ИВФ)</u>...*	30	40
Группа <u>понимать</u>			
«под» T<:case=«ins»> [«мы»] V<«понимать»:tense=«pres» person=«1» number=«plur»>Ng<:case=«acc»>	*Под <u>текстом</u> понимаем множество предложений...*	5	100
Группа <u>быть + понимать</u>			
«под» T<:case=«ins»> [«здесь»] V<«быть»:tense=«fut» person=«1» number=«plur»> «понимать» Ng<:case=«acc»>	*Под <u>объектом</u> здесь будем понимать любые ...*	3,1	100
Группа <u>пониматься</u>			
«под» T<:case=«ins»> V<«пониматься»:tense=«pres» person=«3»> Ng<:case=«nom»><T.number=V.number>	*Под <u>графемной конструкцией</u> понимается графическая форма...*	13,3	89
Группа <u>далее</u>			
«(далее» [«–»] T<:case= «nom»> «)»	*...все концепты области-источника (далее <u>ОИ</u>)...*	7,3	75