# Терминологический анализ текста на основе лексико-синтаксических шаблонов

# Analysis of text terminology based on lexicosyntactic patterns

**Ефремова Н. Э.** (nvasil@list.ru),
**Большакова Е. И.** (bolsh@cs.msu.ru),
**Носков А. А.** (alexey.noskov@gmail.com),
**Антонов В. Ю.** (avadim@gmail.com)

МГУ им. М. В. Ломоносова, факультет ВМиК

Характеризуются лексико-синтаксические шаблоны, специфицирующие особенности терминоупотребления в научно-технических текстах на русском языке. Описываются результаты экспериментального исследования основанных на этих шаблонах процедур автоматического выявления терминов в текстах.

## 1. Введение

Проблема автоматического выявления в тексте на естественном языке терминов — слов и словосочетаний, называющих понятия определенной проблемной области (ПО), изучается с точки зрения разных приложений, таких как реферирование и аннотирование тестов, извлечение знаний из текстовых источников, создание терминологических словарей, тезаурусов и онтологий ПО. Общим при решении этой проблемы является применение частичного синтаксического анализа текста и опора при распознавании терминов на лингвистические и статистические критерии [1, 2]. Статистические критерии так или иначе основаны на частоте употребления терминов и дают приемлемую точность и полноту распознавания только на корпусах текстов. Лингвистические критерии учитывают в первую очередь типичную структуру именных терминологических словосочетаний (частеречную принадлежность слов, входящих в термин, и синтаксические связи между ними), и реже — контексты употребления терминов, свойственные конкретной узкой ПО. Ряд исследований, проведенных для распознавания терминологических словосочетаний английского и французского языков (в частности, [3]) показывает, что более полное использование информации о лексических и синтаксических особенностях терминоупотреблений в обрабатываемых текстах повышает точность и полноту автоматического распознавания терминов.

Терминологический анализ текста, необходимый в таких приложениях, как литературно-научное редактирование и перевод текстов с одного языка на другой, предполагает распознавание в тексте различных терминоупотреблений, причем как можно более полное (при приемлемой точности). В настоящей работе описывается применяемый для решения этой задачи единый подход к представлению необходимой разнородной лингвистической информации в виде *лексико-синтаксических шаблонов* языка LSPL. Этот язык был предложен в работе [4], и для него были реализованы программные средства автоматического распознавания в тексте на русском языке конструкций по заданному шаблону [5].

Поскольку наиболее характерными в плане терминологии являются тексты научно-технического стиля, именно к ним в первую очередь и применялся указанный подход. В результате изучения различных терминологических словарей (более 15 000 словарных статей) и научно-технических текстов на русском языке (около 330 текстов) из области физики и информатики была проведена формализация лексико-синтаксической информации, необходимой для распознавания терминоупотреблений, и создан представительный набор LSPL-шаблонов. Кроме шаблонов, описывающих структуру терминологических словосочетаний, в набор входят шаблоны характерных конструкций определения новых, *авторских* терминов (например: *Под конвейерным режимом понимают такой вид обработки…*). Поскольку часто анализ текста проис-

ходит с привлечением терминологического словаря, зафиксированные в нем *словарные* термины также представляются в виде шаблонов. Кроме того, набор включает шаблоны *текстовых вариантов* терминов [6], учитывающих возможное варьирование в тексте одного термина (*текстовая коллекция — коллекция текстов — коллекция*) и **соединение** в тексте нескольких терминологических словосочетаний, при которых соединяемые сочетания могут разрываться (*входные и выходные данные — входные данные* и *выходные данные*).

В работе кратко рассматриваются средства языка LSPL, необходимые для формализации терминоупотреблений. Характеризуется созданный набор LSPL-шаблонов и базирующиеся на нем процедуры автоматического выделения в научно-техническом тексте различных терминоупотреблений, в том числе — разрывных. Приводятся результаты экспериментального тестирования этих процедур, и формулируется стратегия их совместного применения, позволяющая в целом улучшить показатели распознавания терминов в тексте (F-меру).

## 2. Формализация терминоупотреблений на основе языка LSPL

В общем случае лексико-синтаксический шаблон языка LSPL задает последовательность *элементов-слов*, из которых должна состоять описываемая языковая конструкция, и указывает условия синтаксического согласования этих элементов. Например, шаблон `A N <A=N>` описывает словосочетания из прилагательного `A` и существительного `N`, согласованных по их общим морфологическим характеристикам: *логический вывод*, *реактивной силы*. Для элементов-слов могут быть заданы не только часть речи, но и конкретизированы их отдельные морфологические характеристики (род, число, падеж и др.), а также — лексема, например, шаблон `N <базис,N=sing>` описывает все формы единственного числа слова *базис*.

Для описания конкретных словоформ и символов, встречающихся в описываемой конструкции, в шаблоне используется *элемент-строка*, например: `"базисом"`.

**Язык шаблонов является достаточно мощным, в шаблоне могут задаваться такие сложные элементы, как:**

- *повторения элементов*, задаваемые в фигурных скобках с указанием количества повторений;
- *опциональные элементы*, записываемые в квадратных скобках;
- *альтернативные элементы*, указываемые через символ |.

Например, шаблон `{A} N1 [N2<c=gen>] <A=N1>` определяет именную группу из нескольких прилагательных, согласованного с ними существительного и опционального существительного в родительном падеже (*немаркированный квантор общности, двойной электрический слой*). Заметим, что для обозначения в шаблоне разных элементов одной части речи (существительных `N1` и `N2`) используются числовые индексы.

Язык предоставляет возможность давать имена шаблонам и устанавливать у шаблонов параметры, что позволяет использовать для описания сложных конструкций уже определенные шаблоны (с конкретизацией при необходимости морфологических характеристик входящих в них слов-элементов). Например, задав имя `NP` для вышерассмотренного шаблона и установив его параметрами морфологические характеристики входящего в него существительного:

```
NP = {A} N1 [N2<c=gen>] <A=N1> (N1),
```

можно описать шаблон именной группы, стоящей в творительном падеже: `NP<c=ins>`.

Дополнительно, кроме условий согласования элементов, в шаблоне могут быть заданы *словарные условия*: запись `<Syn(N2,N4)>` означает, что существительные `N2` и `N4` являются синонимами, зафиксированными в соответствующем словаре.

Рассмотренные средства языка шаблонов LSPL позволили описать:

1. морфосинтаксические образцы терминологических слов и словосочетаний;
2. типичные конструкции определения авторских терминов;
3. характерные контексты введения синонимов терминов;
4. входы используемого при распознавании терминов терминологического словаря;
5. правила образования лексико-синтаксических вариантов терминов;
6. правила образования в тексте соединений нескольких терминологических словосочетаний.

Примеры шаблонов каждой группы представлены в Табл. 1; ее последний столбец содержит соответствующие примеры терминов и их употреблений.

В первую группу шаблонов включены 7 наиболее частотных морфосинтаксических образцов терминологических слов и двух- и трехсловных сочетаний. Каждый шаблон фиксирует часть речи входящих в термин слов и их морфологические характеристики.

Вторая группа шаблонов получена в ходе формализации характерных для научной прозы конструкций, используемых при введении в текст новых, *авторских* терминов, например: *Слабовзаимодействующие массивные частицы назовем **вимпами***. В шаблонах этой группы используется вспомогательный шаблон с именем `Term`, описывающий возможные морфосинтаксические образцы определяемого термина, и шаблон `Defin`, задающий структуру определяющей термин конструкции.

**Таблица 1.** Лексико-синтаксические шаблоны

| № | Группы шаблонов | Примеры шаблонов | Примеры терминов и их употреблений |
|---|---|---|---|
| 1 | Морфо-синтаксические образцы терминов | N1 (N1) | *вимп* |
| | | A1 N1 <A1=N1> (N1) | *опорная точка* |
| | | N1 N2<c=gen> (N1) | *период упреждения* |
| | | N1 A2 N2<c=gen> <A2=N2> (N1) | *технология двойной накачки* |
| 2 | Контексты определения авторских терминов | Defin<c=acc> "будем" "называть" Term<c=ins> #Term<c=nom> | *Такие операции будем называть понятийными операциями* |
| | | "под" Term<c=ins> "понимается" Defin<c=nom> #Term<c=nom> | *Под прерыванием понимается сигнал…* |
| 3 | Контексты введения синонимов терминов | Term1 "("Term2")" <Term1.c=Term2.c> #Term1<c=nom>, Term2<c=nom> | *взаимодействующих компонентов (подсистем)* |
| | | Term1 "," "или" Term2 <Term1.c=Term2.c> #Term1<c=nom>, Term2<c=nom> | *разрядностью, или длиной слова* |
| 4 | Словарные термины | N1<вектор> [N2<намагниченности,c=gen>\| N2<состояния,c=gen>\|"Умова"] | *вектор*, *вектор намагниченности*, *вектор состояния*, *вектор Умова* |
| | | A1<битовый> {N2<массив>\| N2<образ>}<1,1> <A1=N2> | *битовый массив*, *битовый образ* |
| 5 | Лексико-синтаксические варианты | N1 N2<c=gen> #N1, N1 N4<c=gen> <Syn(N2,N4)>, N3 N2<c=gen> <Syn(N1,N3)>, A1 N1 <A1.st=N2.st> | *вывод данных — вывод* (N1), *вывод информации* (N1 N4) |
| | | | *шина адреса — шина* (N1), *адресная шина* (A1 N1) |
| 6 | Соединения терминов | "как" A1 "," "так" "и" A2 N1 <A1=A2=N1> #A1 N1, A2 N1 | *как тонкий, так и толстый клиент — тонкий клиент*, *толстый клиент* |
| | | N1 N2<c=gen> "," N3<c=gen> {"и"\|"или"} N4<c=gen> #N1 N2<c=gen>,N1 N3<c=gen>, N1 N4<c=gen> | *шинам адреса, данных и управления — шина адреса*, *шина данных*, *шина управления* |
| | | A1 A2 N1 <A1=A2=N1> #A1 N1, A2 N1 | *удаленный банковский терминал — банковский терминал*, *удаленный терминал* |
| | | N1 A2 N2<c=gen> <A2=N2> #N1 N2, A2 N2 | *разрядность внутренних регистров — разрядность регистра*, *внутренний регистр* |

Заметим, что все шаблоны второй группы включают в свой состав элемент #Term <c=nom> — выделяемую конструкцию. В общем случае *выделяемая конструкция* (записывается после знака #) определяет, какая часть распознанной конструкции должна быть из нее выделена и с какими морфологическими характеристиками. В данном случае задается выделение распознанного авторского термина, причем в нормализованной форме — в именительном падеже (для рассмотренного примера — *вимп*).

Третья группа шаблонов строилась аналогично второй — в нее входят шаблоны типичных контекстов, в которых вводятся синонимы терминов, например: *…проектирование прикладного ПО (приложений)*. В шаблонах задается выделение пары синонимичных терминов (*прикладное ПО* и *приложения*).

Язык лексико-синтаксических шаблонов оказался удобным и для записи входов терминологических словарей (четвертая группа шаблонов). Каждый шаблон позволяет описать в общем случае несколько *словарных* терминов, имеющих общее начало.

Последние две группы шаблонов описывают в общем виде текстовые варианты терминов: *лексико-синтаксические варианты* одного термина и *соединения* нескольких терминологических словосочетаний (в Табл. 1 в соответствующих примерах шаблонов обоих групп для краткости опущены условия нормализации).

Шаблоны пятой группы по сути формализуют правила образования вариантов термина: каждый шаблон фиксирует один из морфосинтаксических образцов термина и задает (после знака #) вы-

деляемые конструкции — возможные текстовые варианты термина (тоже в виде морфосинтаксических образцов). Например, для терминологических словосочетаний со структурой `N1 N2<c=gen>` (см. строку 5 Табл. 1) учтены следующие случаи:

1) Вставка или отбрасывание слова — вариант `N1` (*ввод данных — ввод*).
2) Замена одного слова на синоним в данной ПО — варианты `N1 N4<c=gen>` (*вывод информации — вывод данных*) и `N3 N2<c=gen>` (*метка адреса — маркер адреса*), при этом предполагается выполнение соответствующих словарных условий синонимии.
3) Замена слова на однокоренное другой части речи с одновременным изменением синтаксических связей словосочетания — вариант `A1 N1 <A1=N1>` (*шина адреса — адресная шина*), условие равенства корней слов записывается как `A1.st=N2.st`.

Аналогично были формализованы правила образования типичных *соединений* в тексте нескольких терминологических словосочетаний, при которых один или несколько терминов разрываются и/или усекаются. Среди соединений терминов мы различаем:

- соединения с помощью сочинительных союзов (*шина адреса, данных и управления — шина адреса*, *шина данных*, *шина управления)* и
- бессоюзные соединения (*разрядность внутренних регистров — разрядность регистров*, *внутренние регистры*).

Каждый шаблон соединения задает выделение всех входящих в него терминов.

## 3. Процедуры распознавания и их тестирование

Для каждой из описанных групп лексико-синтаксических шаблонов была разработана процедура автоматического распознавания в научно-техническом тексте соответствующих терминоупотреблений. Эти процедуры позволяют выявлять соответственно термины-кандидаты, авторские термины, термины-синонимы, словарные термины, лексико-синтаксические варианты и соединения терминов. Дополнительно, процедуры подсчитывают частоту употребления каждого из распознанных ими терминов. Результат распознавания терминов по их морфосинтаксическому образцу назван нами *терминами-кандидатами*, чтобы подчеркнуть, что в их числе с достаточно большой вероятностью могут оказаться общенаучные словосочетания вида *решение задачи*, *применение последнего правила*, не являющиеся терминами.

**На вход** каждой процедуре, за исключением процедуры выявления лексико-синтаксических вариантов, поступает анализируемый текст и соответствующая группа шаблонов. На первом этапе работы процедуры выполняется поиск всех фрагментов текста, представляющих собой искомые терминоупотребления, а на втором — подсчитывается частота употребления каждого выявленного в них термина. Разделение этих двух этапов необходимо для корректного подсчета частоты употребления в случаях полного вложения одного термина в другой (*адрес — логический адрес*), поскольку частота должна быть определена без учета таких вложений. **Таким образом, на выходе** указанных процедур получается список фрагментов-терминоупотреблений и список выделенных терминов с частотой их употребления в тексте.

В процедурах, опирающихся на шаблоны с выделяемыми конструкциями (шаблоны авторских терминов, синонимов и соединений), этап выявления происходит в два приема. Сначала в тесте ищутся все фрагменты, соответствующие шаблонам и из них выделяются термины, точнее шаблоны, их описывающие, и эти шаблоны используются затем для поиска всех употреблений этих терминов. В частности, при работе процедуры выявления авторских терминов сначала ищутся контексты определения терминов, из них выделяются авторские термины, вхождения которых потом ищутся в тексте.

По иному организована процедура выявления лексико-синтаксических вариантов терминов. **На вход** ей поступают:

- шаблоны правил образования лексико-синтаксических вариантов;
- термины, для которых необходимо найти их варианты;
- слова и словосочетания, среди которых необходимо искать эти варианты.

**На выходе** процедуры получаются группы эквивалентных вариантов; каждая группа объединяет слова и словосочетания, соответствующие (предположительно) одному и тому же понятию ПО.

Все разработанные процедуры были по отдельности протестированы на научно-технических текстах из разных областей физики и информатики (общего объема 700 Кбайт). При этом для выявления словарных терминов использовались словари по физике (более 3 тыс. терминов) и по информатике (более 4 тыс. терминов). Для выявления лексико-синтаксических вариантов был создан рабочий словарь терминологических синонимов в этих ПО.

Примерно для трети текстов результаты работы процедур были сравнены со списками терминов, выявленными в текстах экспертами, при этом оценивались полнота и точность как выделения самих терминов, так и их употреблений в тексте — см. Табл. 2. Для процедур распознавания синонимов и соединений замерялась полнота и точность выделения терминов, встретившихся в рамках распознаваемых ими конструкций.

**Таблица 2.** Полнота и точность процедур

| Процедура | Выделение терминов | | Выделение терминоупотреблений | |
|---|---|---|---|---|
| | Полнота | Точность | Полнота | Точность |
| Термины-кандидаты | 58 % | 24 % | 54 % | 25 % |
| Авторские термины | 67 % | 89 % | 70 % | 97 % |
| Синонимы | 57 % | 22 % | – | – |
| Словарные термины | 85 % | 94 % | 87 % | 95 % |
| Соединения | 71 % | 30 % | – | – |

Наихудшие результаты (что было вполне прогнозируемо) дала процедура выявления терминов-кандидатов, опирающаяся на минимум лингвистической информации. Она давала много «шума» (*крупный размер, аналогичный результат*), в тоже время не были *распознаны сочетания, чья морфосинтаксическая структура не учтена в наборе шаблонов (например, термины вида индекс iCOMP и обратная связь по релевантности)*.

Для **словарных терминов** нераспознанными оказались преимущественно их терминоупотребления внутри соединений, *а некоторые распознанные термины* оказались частью несловарных терминов (например, термин *ряд* — частью общенаучных выражений: *в ряде случаев*, *за рядом исключений)*.

Для **авторских терминов** и синонимов основная потеря полноты возникла опять же из-за ограниченности морфосинтаксических образцов терминов, но также и по причине неучета некоторых конструкций определения терминов и их синонимов (к примеру, *Регистр представляет собой совокупность бистабильных устройств*). Подобные конструкции рассматривались нами при построении набора шаблонов, но не были включены в него, поскольку часто имеют смысл, отличный от определения. Неточность же выявления терминов касалась тех случаев, когда в характерном контексте определения термина уточнялось значение словарных терминов.

Что касается процедур выявления соединений, то причины невысоких значений полноты и точности такие же, как и для группы терминов-кандидатов.

## 4. Стратегия совместного применения процедур

В целом, проведенный анализ результатов работы процедур показал, во-первых, что полноту распознавания можно повысить, расширяя набор шаблонов, но при этом часто страдает точность. Во-вторых, возможный учет процедурами результатов работы других процедур в ряде случаев может повысить полноту распознавания: так, в случае учета выявленных из соединений разрывных терминов наблюдался прирост полноты в среднем на 12 %. В-третьих, использование результатов других процедур позволяет более точно определять частоту употребления каждого конкретного термина в тексте.

Поскольку простое объединение результатов работы процедур (вырабатываемых ими списков терминов) повышает полноту выявления терминов, достаточно сильно снижая точность, были изучены другие способы объединения. Это позволило сформулировать соответствующую стратегию совместного применения процедур, согласно которой:

1) К заданному тексту применяются все процедуры распознавания, за исключением процедуры поиска лексико-синтаксических вариантов.
2) Распознанные словарные и авторские термины, объединяются и включаются в формируемое множество выделенных терминов, причем в случае полных вложений предпочтение отдается более длинным терминам.
3) Из выявленных терминов-кандидатов в полученное множество включаются только те, в состав которых входят словарные или авторские термины, при этом последние исключаются из формируемого множества.
4) Из найденных пар терминологических синонимов берутся только пары, один член которых уже входит в формируемое множество, и в него добавляется второй член пары.
5) Термины из выявленных соединений добавляются во множество, только если среди них есть разрывный словарный термин.
6) Для сформированного множества терминов применяется процедура поиска их лексико-синтаксических вариантов из числа оставшихся терминов-кандидатов, и полученные варианты добавляются к формируемому множеству.
7) Во множество дополнительно добавляются термины из тех соединений, в которые входят текстовые варианты, найденные на предыдущем этапе.
8) Из оставшихся терминов-кандидатов в формируемое множество добавляются те, частота употребления которых выше заранее установленного порога (например, равного среднему квадратичному частот терминов-кандидатов).

Поскольку показатели полноты и точности распознавания взаимосвязаны (как правило, увеличение одного показателя приводит к уменьшению другого) для оценки результатов рассмотренной стратегии использовалась F-мера — комбинированный

показатель, вычисляемый как гармоническое среднее полноты и точности. Оказалось, что в большинстве случаев применение стратегии дает ощутимое повышение F-меры по сравнению с простым объединением результатов работы процедур распознавания. К примеру, для нижеследующего текста был зафиксирован прирост F-меры выявления терминов на 19,8 %, а F-меры выявления терминоупотреблений — на 15,5 % (все выявленные терминоупотребления подчеркнуты: двойной линией подчеркнуты слова, являющиеся частями нескольких терминов, пунктирной линией — слова, являющиеся частями разрывных терминов):

*Микропроцессор, как правило, представляет собой сверхбольшую интегральную схему, реализованную в едином полупроводниковом кристалле и способную выполнять функции центрального процессора. С внешними устройствами микропроцессор может «общаться» благодаря шинам адреса, данных и управления, выведенным на специальные контакты корпуса микросхемы. Стоит отметить, что разрядность внутренних регистров микропроцессора может не совпадать с количеством внешних выводов для линий данных. Иначе говоря, микропроцессор с 32-разрядными регистрами может иметь, например, только 16 линий внешних данных.*

*Любое внешнее устройство, совершающее по отношению к микропроцессору операции ввода-вывода, можно назвать периферийным.*

*Порт — это некая схема сопряжения, обычно включающая в себя один или несколько регистров ввода-вывода и позволяющая подключить, например, периферийное устройство к внешним шинам микропроцессора. Практически каждая микросхема использует для различных целей несколько портов ввода-вывода.*

В приведенном тексте несколько терминов остались нераспознанными (например, *контакты корпуса* — по причине отсутствия в словаре по информатике), в то же время выявлено как термин общенаучное словосочетание *различные цели*. Дальнейшая настройка стратегии требует проведения дополнительных экспериментов.

## 5. Заключение

Разработаны процедуры выявления терминоупотреблений в заданном научно-техническом тексте на основе набора лексико-синтаксических шаблонов, полученных в ходе формализации различных случаев употребления в текстах терминологических слов и словосочетаний. На основе анализа результатов их раздельного тестирования предложена стратегия совместного применения этих процедур, позволяющая повысить F-меру полноты и точности распознавания. Для научной прозы дальнейшие резервы повышения полноты связаны с учетом дополнительных видов текстовых вариантов, а точности — с учетом взаимного расположения вариантов в тексте и словаря общенаучных выражений.

Поскольку используемый набор шаблонов может быть изменен без перепрограммирования процедур распознавания, это дает возможность настройки процедур на обработку текстов разных ПО с учетом присущих им особенностей терминоупотребления.

## Литература


1. *Jacquemin C., Bourigault D.* Term extraction and automatic indexing // Mitkov R. (ed.): Handbook of Computational Linguistics. Oxford University Press, 2003. P. 599–615.
2. *Добров Б. В., Лукашевич Н. В., Сыромятников С. В.* Формирование базы терминологических словосочетаний по текстам предметной области // Труды пятой всероссийской научной конференции «Электронные библиотеки: Перспективные методы и технологии, электронные коллекции». 2003. С. 201–210.
3. *Nenadic G., Ananiadou S., McNaught J.* Enhancing Automatic Term Recognition through Variation // Proceedings of 20th Int. Conference on Computational Linguistics COLING'04. 2004. P. 604–610.
4. *Большакова Е. И., Баева Н. В., Бордаченкова Е. А., Васильева Н. Э., Морозов С. С.* Лексико-синтаксические шаблоны в задачах автоматической обработки текстов // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог '2007. М.: Изд-во РГГУ, 2007. С. 70–75.
5. *Носков А. А.* Метод выделения в тексте конструкций по их лексико-синтаксическим шаблонам // Сборник статей молодых ученых факультета ВМиК МГУ. М.: Издательский отдел фак-та ВМиК МГУ им. М. В. Ломоносова; МАКС Пресс, 2009. Выпуск 6. С. 136–145.
6. *Большакова Е. И., Васильева Н. Э.* Терминологическая вариантность и ее учет при автоматической обработке текстов // Одиннадцатая национальная конференция по искусственному интеллекту с международным участием: Труды конференции. М.: ЛЕНАНД, 2008. Т. 2. С. 174–182.