УДК 681.3

ЛЕКСИКО-СИНТАКСИЧЕСКИЕ ШАБЛОНЫ ДЛЯ АВТОМАТИЧЕСКОГО АНАЛИЗА НАУЧНО-ТЕХНИЧЕСКИХ ТЕКСТОВ*

Е.И. Большакова[1], Н.Э. Васильева[2], С.С. Морозов[3]

Рассматриваются лексико-фразеологические и дискурсивные особенности текстов научно-технического стиля, которые следует учесть при разработке процедур автоматической обработки текстов. Характеризуются разрабатываемые словарные средства, отражающие указанные особенности: компьютерный словарь общенаучной речи и лексико-синтаксические шаблоны типичных фраз. Кратко описываются составные элементы и язык записи шаблонов, а также методика их разработки.

Введение

Одним из наиболее своеобразных функциональных стилей речи является научный стиль, представленный в текстах из различных научно-технических областей естественных и точных наук. Научную прозу отличает не только ее высокая стандартизированность и насыщенность специальными терминами, но и особый формально-логический способ изложения материала. Типичное научное изложение представляет собой рассуждение, призванное описать и обосновать результаты проведенного научного исследования. Шаги рассуждения обычно указываются автором текста при помощи общенаучных слов и выражений (слов-организаторов научной мысли): *далее мы докажем*, *из вышесказанного следует*, *в заключение*, *по причине того, что* и т.п. Такие слова и выражения называются также дискурсивными (речевыми) маркерами, поскольку они помечают дискурсивные операции и относятся к дискурсивному уровню текста. Из общенаучных слов строятся типичные выражения-клише

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

[1] *119992, Москва, ГСП-2, Ленинские горы, МГУ им. М. В. Ломоносова, Факультет ВМиК, bolsh@cs.msu.su*

[2] *119992, Москва, ГСП-2, Ленинские горы, МГУ им. М. В. Ломоносова, Факультет ВМиК, nvasil@list.ru*

3 119992, Москва, ГСП-2, Ленинские горы, МГУ им. М. В. Ломоносова, Факультет ВМиК, sergej_morozov@rambler.ru

научной прозы: *как показало проведенное исследование*, *всесторонний анализ проблемы приводит к выводу* и т.п.

Лексико-фразеологические и дискурсивные особенности текстов научно-технического стиля взаимосвязаны и носят системный характер, и они должны быть всесторонне учтены при разработке процедур, автоматизирующих отдельные интеллектуальные операции над текстом. Для выявления этих особенностей были изучены общенаучная фразеология и дискурсивная организация научных текстов разных жанров (статьи, монографии, аннотации, справочники) из разных предметных областей; преимущественно рассматривались научные статьи, как относящиеся к «ядру» функционального стиля.

Выявленные особенности легли в основу рабочей гипотезы, согласно которой процедура автоматического распознавания общей дискурсивной структуры научного текста и примененных в нем операций научного мышления (по сути – поверхностного его понимания) может быть построена на основе частичного синтаксического анализа текста и лексикона общенаучных слов и выражений, без использования полного синтаксического разбора предложений. В работе [Севбо, 1989] аналогичная идея выдвигалась и обсуждалась применительно к текстам любого стиля, но не была воплощена в работающей системе.

В ходе наших исследований была начата разработка словарных средств, отображающих специфику именно научной прозы: компьютерного словаря общенаучных слов и выражений и специальных лексико-синтаксических шаблонов, описывающих характерные конструкции научно-технических текстов [Большакова и др., 2004; Васильева, 2004]. Были определены составные элементы лексико-синтаксических шаблонов, язык их записи и методика создания набора шаблонов, на основе который был разработан первоначальный набор шаблонов для описания характерных конструкций определения новых терминов. Данные лексико-синтаксические шаблоны подобны тем, что предложены в работе [Hearst, 1998], но используются для других целей: на набор лексико-синтаксических шаблонов опирается разрабатываемая нами процедура распознавания дискурсивной структуры научного текста.

К числу прикладных задач, для решения которых требуются разрабатываемые процедуры и словарные средства, относятся:

- Литературно-научное редактирование и обучение научно-техническому литераторству;
- Реферирование и аннотирование научно-технических текстов;
- Извлечение знаний из научных текстов, в том числе – определений новых понятий;
- Структуризация научно-технических текстов для быстрого внутритекстового поиска.

1. Научный дискурс и общенаучный лексикон

Основная цель научного произведения – сообщение о результатах проведенного исследования и объяснение способа их получения, формулировка новых идей и их обоснование. Соответственно, научный дискурс представляет собой логически взаимосвязанную последовательность речевых (дискурсивных) действий, соответствующих операциям научного мышления [Митрофанова, 1973; Николаев, 1998; Рябцева, 1992]. К типичным операциям относится обоснование вывода, выдвижение гипотезы, введение термина и понятия, приведение фактов и доказательств, подведение итогов и др. Как правило, эти операции более или менее явно помечаются общенаучными словами и выражениями, образующими общенаучный лексикон.

Наиболее явными маркерами мыслительных операций служат так называемые ментальные перформативные высказывания (например: *ниже рассмотрим*, *особо подчеркнем*), которые обычно квалифицируют применяемую операцию. В работе [Рябцева, 1992] описаны виды перформативных высказываний, опирающиеся на широкий круг ментальных перформативных глаголов (*опишем*, *предположим*, *заметим* и т.п.):

- канонические, с глаголом в 1 лице множественного числа (*мы покажем*);
- «установочные», с модальным или оценочным словом (*необходимо*/*нетрудно заметить*);
- в форме деепричастия или деепричастного оборота (*резюмируя вышесказанное*);
- в безличной форме (*представляется, что…*).

В научных текстах встречаются также дескриптивные (косвенные) варианты ментальных перформативов, используемые либо для перифразирования (*эти данные приводятся в таблице 3* вместо канонического *мы приводим эти данные в таблице 3*), либо для установления связей между высказываниями текста (*далее кратко изложен*).

Кроме перформативов используются также маркеры очередности (*во-первых, наконец* и др.); коннекторы – союзы и союзные слова (*однако, благодаря тому, что* и т.п.); слова-оценки (*возможно*, *по-видимому* и т.п.), часто встречающиеся и в текстах других стилей [Севбо, 1989].

Все указанные виды дискурсивных маркеров имеют ярко выраженный метатекстовый характер [Вежбицка, 1978], большинство из них функционируют в тексте как метатекстовые операторы, предполагающие в своем составе сентенциальный или атрибутивный аргумент: *подчеркивается, что S*, *рассмотрим N*).

К общенаучному лексикону относятся абстрактные существительные, называющие аппарат научно-познавательной деятельности (*вопрос*, *проблема*, *понятие*, *анализ*, *процедура*, *схема* и др.). Эти существительные называются общенаучными переменными [Севбо, 1989], поскольку имеют обязательную атрибутивную валентность (*проблема N*, *понятие N*). Хотя они не используются в метатекстовой функции, они играют важную роль в структуризации научной информации. Общенаучные переменные обычно употребляются в научных текстах с перформативными глаголами (*ввести понятие*, *подвергнуть анализу*) [Словарь, 1973].

Таким образом, общенаучный лексикон охватывает широкий круг семантически и грамматически разнородных слов и выражений общенаучной речи. Важно, что он не зависит от конкретной предметной области и сравнительно немногочисленен. Заметим, что лексикон состоит из общеупотребительных слов, поэтому их метатекстовая функция в конкретном предложении текста (т.е. выполняют ли они роль дискурсивного маркера) в общем случае может быть установлена только в результате исследования контекста их употребления.

2. Словарные средства анализа научно-технических текстов

Для анализа научно-технических документов предлагается использовать, кроме традиционных терминологического и морфологического словарей, следующие словарные средства, отображающие специфику научной прозы:

- Словарь общенаучных слов и выражений;
- Лексико-синтаксические шаблоны типичных фраз научной речи.

При построении словаря общенаучных слов и выражений была проведена функционально-семантическая классификация собранных единиц. Выражения были разбиты на группы исключительно по их смыслу и функции в тексте, без учета их грамматической формы и синтаксических характеристик. В итоге получилось 53 группы, каждая из которых является классом слабой эквивалентности и включает в общем случае несколько семантически близких выражений разной грамматической природы. Каждой группе приписана соответствующая операция научного дискурса; эти операции частично приведены в таблице 1.

Для каждой единицы словаря общенаучной речи указывается ее классификационная группа (дискурсивная операция); для словосочетаний описываются их синтаксические характеристики (разрывность/ неразрывность и др.).

Для распознавания в тексте словарного словосочетания необходима информация о семантико-синтаксических валентностях составляющих его

слов. Такую информацию можно представить в виде лексико-синтаксического шаблона, который фиксирует лексемы и их грамматическую форму, а также задает синтаксические условия заполнения своих пустых мест (валентностей). Кроме того, в виде шаблонов удобно представлять клишированные конструкции научной речи, составленные из нескольких словарных единиц и имеющие фиксированную синтаксическую структуру. К числу таких конструкций относятся определения новых терминов, состоящие из одного предложения, например, фраза *«…значение, которое используется для расширения первоначального набора, мы будем называть существенным значением…»*. Указанная конструкция схематически может быть описана как

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

где «мы» и «будем называть»совместно встречающиеся лексемы, причем слово «мы» может отсутствовать; $\mathbf{T_{INS}}$ – определяемый термин, выраженный согласованной именной группой, главное слово которой имеет форму творительного падежа; $\mathbf{NG_{ACC}}$ – определение или объяснение авторского термина, выраженное согласованной именной группой (возможно, расширенной придаточным предложением), главное слово которой имеет форму винительного падежа.

Таблица 1. Операции научного дискурса

Операции	**Примеры слов и выражений**
Описание и констатация	*укажем, что; характеризуя*
Конкретизация и уточнение	*в частности; в дополнение к*
Причинно-следственные связи	*по этой причине; следовательно*
Актуализация темы	*перейдем к; рассмотрим*
Выделение информации	*особо подчеркнем; необходимо отметить*
Предположения и допущения	*предположим/допустим, что*
Определения	*будем называть; по определению*
Сравнение и противопоставление	*с одной стороны; в отличие от; по сравнению с*
Иллюстрация и приведение примеров	*к примеру; например*
Обобщение и резюмирование	*суммируя вышесказанное; в общем*
Упорядочивание и перечисление	*во-первых; наконец*
Помета общенаучной переменной	*идея, модель, результат*
Выражение мнения и оценивание	*целесообразно считать; по-видимому*

Представленная в словаре общенаучной речи и наборе шаблонов семантико-синтаксическая информация позволяет производить содержательный анализ научно-технических текстов – распознавание примененных дискурсивных маркеров и операций научного дискурса.

3. Разработка лексико-синтаксических шаблонов

Проблема, возникающая при разработке шаблонов конструкций, заключается в определении контекстов, однозначно сигнализирующих дискурсивный (метатекстовый) характер употребляемых слов и словосочетаний. Для ее решения необходимо проводить исследование контекстов употреблений конструкций.

Такое исследование было проведено для контекстов конструкций, определяющих новые термины. Вручную было просмотрено около 50 научно-технических текстов, и из них были выделены те фразы, которые использовались при определении или пояснении нового термина. После их предварительного анализа было получено первоначальное множество лексем, входящих в конструкции определений, что позволило в дальнейшем частично автоматизировать процесс поиска новых конструкций и контекстов.

Так как количество разных контекстов было велико, контексты для каждой фиксированной лексемы (или для двух-трех совместно встречающихся лексем) были рассмотрены по отдельности, что позволило выявить соответствующие синтаксические конструкции, которые затем были формализованы в виде лексико-синтаксических шаблонов.

В состав шаблонов входят следующие элементы:

- Литералы, т.е. конкретные лексемы из словаря общенаучной речи (**«определим», «будем называть»** и др.), а также сокращения (**«т.н.»**) и знаки препинания. Литеральные элементы заключаются в кавычки.
- Символьные обозначения слов определенной части речи и грамматической формы, которые могут заполнять свободные места (слоты) шаблона; например, **N** – существительное, **V** – глагол, **P** – предлог, **Pa** – причастие.
- Символьные обозначения определенных грамматических конструкций, например, **Ng** – именная группа, **T** – определяемый термин, выраженный именной группой (простой или расширенной).
- Условия, уточняющие грамматические характеристики рассмотренных элементов и записываемые в угловых скобках, например: **<Ng.number=V.number>** означает, что число группы **Ng** и глагола **V** совпадают, а условие **<person=3>** фиксирует употребление третьего лица.

При записи условий используются символьные обозначения грамматических характеристик: времени (**tense**), числа (**number**), лица

(**person**), рода (**gender**), падежа (**case**) и конкретных падежей (например, **nom** – именительный, **ins** – творительный).

К примеру, шаблон

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

описывает случаи вида *«По результатам генерации форм, слова были разбиты на группы, названные профилями»* (в этом примере подчеркнута фиксированная шаблоном лексема). 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 стоит конструкция, имеющая требуемые в шаблоне характеристики.

Разработанными к настоящему моменту шаблонами покрывается примерно 60-70% процентов определений терминов, встречающихся в научных текстах. Важно, что добавляя новые шаблоны, учитывающие все более сложные конструкции и контексты, можно постепенно наращивать мощность процедуры распознавания в тексте операций научного дискурса.

4. Применение шаблонов для автоматического анализа текста

Процедура автоматического анализа текста, последовательно обрабатывающая его предложения и использующая описанные выше словарные средства, включает среди прочих следующие шаги:

- Выделение слов и словосочетаний общенаучной лексики. Например, во фрагменте *«Таким образом, актуальной является задача разработки…»* будут распознаны следующие общенаучные словосочетания: *таким образом*, *являться актуальной задачей*.
- Наложение лексико-семантических шаблонов и распознавание дискурсивных маркеров. Если в предложении встретилось слово, являющееся литералом некоторого шаблона, то происходит сопоставление предложения с этим шаблоном, при этом проверяются синтаксические условия для заполняемых мест шаблона. В случае успешного сопоставления происходит заполнение слотов шаблона (что фактически означает извлечение из анализируемого предложения языковых конструкций).

При этом полный синтаксический разбор предложений не производится; в то же время при необходимости осуществляется проверка согласования и управления слов (например, согласование составляющих слов в употребленных словарных выражениях).

Рассмотрим шаблон:

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

Он 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».

В результате успешного сопоставления вышеописанного шаблона с первым из приведенных определений будет выделен новый термин *графемная конструкция* и определяющая его конструкция – см. Рисунок 1.

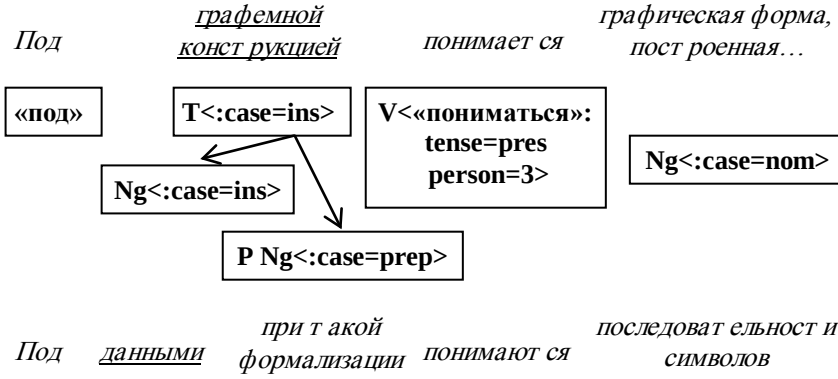

Рисунок 1. Схема применения шаблона

При сопоставлении шаблона со вторым из приведенных определений будет учтено, что в качестве заполнителя его слота **T** может быть взята расширенная именная группа (т.е. простая именная группа, за которой следует предложная именная группа, играющая роль обстоятельства), и в результате сопоставления будет правильно выделен новый термин *данные* и определяющая его конструкция.

Заключение

Описаны основные особенности дискурса и лексикона научно-технических текстов, учет которых позволяет организовать их содержательный анализ без глубокого синтактико-семантического разбора предложений текста. Охарактеризованы разрабатываемые для этого

словарные средства – словарь общенаучной речи и лексико-синтаксические шаблоны характерных фраз. Кратко описаны составные элементы шаблонов, язык их записи, а также методика их построения, базирующаяся на анализе контекстов. Все это дает возможность приступить к реализации процедуры распознавания дискурсивной структуры научно-технических текстов.

Список литературы

[Большакова и др., 2004] Большакова Е.И., Баева Н.В., Васильева Н.Э. Структурирование и извлечение знаний, представленных в научных текстах // Девятая Национальная конференция по искусственному интеллекту с международным участием КИИ-2004. Труды конференции в 3-х томах. Т. 2. М.: Физматлит, 2004.

[Васильева, 2004] 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 / Под ред. И.М. Кобозевой, А.С. Нариньяни, В.П.Селегея. М.: Наука, 2004.

[Вежбицка, 1978] Вежбицка А. Метатекст в тексте // Новое в зарубежной лингвистике. Вып. VIII. М.: Прогресс, 1978.

[Митрофанова, 1973] Митрофанова О.Д. Язык научно-технической литературы. – М.: Изд-во МГУ, 1973.

[Николаев , 1998] Николаев А.М. Описание семантики научного текста с позиций теории речевых актов (на материале рецензии на научно-техническую работу) // НТИ. Сер. 2. 1998, № 7.

[Рябцева, 1992] Рябцева Н.К. Ментальные перформативы в научном дискурсе // Вопросы языкознания. 1992, № 4.

[Севбо, 1989] Севбо И.П. Сквозной анализ как шаг к структурированию текста // НТИ. Сер. 2. 1989, № 2.

[Словарь, 1973] Словарь глагольно-именных словосочетаний общенаучной речи. – М., Наука, 1973.

[Hearst, 1998] Hearst, M.A. Automated Discovery of WordNet Relations. In: Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database. MIT Press, Cambridge, 1998.