УДК 681.3

СТРУКТУРИРОВАНИЕ И ИЗВЛЕЧЕНИЕ ЗНАНИЙ, ПРЕДСТАВЛЕННЫХ В НАУЧНЫХ ТЕКСТАХ

Н.В.Баева[1], Е.И.Большакова[2], Н.Э.Васильева[3]

Введение

Научно-техническая проза в ее жанровом многообразии (научные статьи, монографии, аннотации, технические отчеты и др.) выполняет функции оформления, сохранения и передачи научно-технической информации. Еще одной важной функцией научного текста является выработка новых знаний, прежде всего – понятийных. Но воплощенное в научном тексте знание не исчерпывается содержанием терминов (понятий) и их смысловых связей, необходимо рассмотрение различных аспектов этого знания, в том числе – композиционно-речевого.

Особый интерес с этой точки зрения представляет жанр научной статьи, относящийся к «ядру» научного стиля. Научная статья ставит своими целями описание результатов проведенного исследования, объяснение способа их получения, формулировку новых идей и их обоснование. Соответственно, научное изложение состоит главным образом из рассуждений, организованных как логическая последовательность шагов информирования, аргументирования и оценки. Как правило, компоненты рассуждения помечаются (маркируются) общенаучными словами и выражениями (например: *по этой причине*, *суммируя вышесказанное*, *предположим, что* и т.п.). Такие слова и выражения называются также словами-организаторами научной мысли, поскольку основное их назначение – структурно-смысловая организация научного текста, т.е. оформление и упорядочение рассуждений, связывание отдельных текстовых фрагментов. С помощью этих слов и выражений формируется логико-композиционная структура текста, представляющая один из важных пластов научного знания, которое следует извлекать из научных текстов.

[1] *119899, Москва, Воробьевы горы, МГУ им. М.В. Ломоносова, Факультет ВМиК; baeva@vog.ru*

[2] *МГУ им. М.В.Ломоносова, Факультет ВМиК, bolsh@cs.msu.su*

[3] *МГУ им. М.В.Ломоносова, Факультет ВМиК, nvasil@port.ru*

В отличие от большинства работ по тематике приобретения знаний из текстов (например, [Осипов, 1997]), посвященных выявлению системы понятий и их семантических связей, в настоящей работе рассматривается как понятийный, так и композиционно-речевой аспект знаний, представленных в отдельных научных текстах. Нашей целью является разработка автоматизированных процедур, выявляющих логико-композиционную структуру научного текста и новые, введенные автором текста, термины (понятия) и их определения. По существу, эти процедуры призваны реализовать начальный этап извлечения научного знания – структурирование содержания научного текста.

Важной чертой нашего подхода является учет стилевых особенностей научных произведений – как терминологических, так и логико-композиционных. Разрабатываемые процедуры автоматизированного анализа опираются на словарные средства: компьютерный словарь общенаучной речи и лексико-синтаксические шаблоны термино-употреблений, отражающие специфику научно-технической прозы.

Терминологические особенности научного текста

Под терминами обычно понимают устоявшиеся в своем значении слова и словосочетания, обозначающие понятия некоторой проблемной области, принадлежащие сложившейся терминологической системе и зафиксированные в соответствующих терминологических словарях [Митрофанова, 1973]. Такие термины называются словарными [Пшеничная и др., 1991] – для них свойственна четкость значения, а также устойчивость лингвистической формы многословных терминов. Устойчивость означает, что для обозначения одного и того же понятия обычно используется одно терминологическое словосочетание, изредка – несколько синонимичных вариантов: грамматических (*время центрального процессора – процессорное время*) и семантических (*машинная графика – компьютерная графика*), используются также сокращения (*база данных – БД*).

В научно-технических текстах встречаются термины, не являющиеся общепринятыми и отсутствующие в терминологических словарях. Они выражают новые понятия, вырабатываемые в процессе исследовательской деятельности. В большинстве случаев эти термины определяются и подробно характеризуются автором в научном тексте – такие термины мы называем авторскими [Bolshakova, 2001]. Вместе со словарными авторские термины выражают логико-понятийный аппарат научного текста.

По своей лингвистической форме авторские термины, как и словарные, представляют собой имена существительные и именные словосочетания нескольких грамматических образцов [Митрофанова, 1973]. Но в отличие

от словарных, для авторских многословных терминов характерна неустойчивость лингвистической формы: даже в пределах текста, в котором они определяются, могут встречаться два-три и даже более их синонимичных варианта: грамматических и семантических, а также сокращенных словосочетаний (*мультиплексор программно-аппаратного комплекса Internet-канал – мультиплексор Internet-канала*).

Отметим также, что авторские термины могут вводиться в текст и употребляться в нем без формулировки их определений. Именно лингвистическая неустойчивость и возможное отсутствие определений существенно усложняют распознавание в научном тексте авторских терминов и извлечение определяющих (или поясняющих) их языковых выражений. Для распознавания авторских терминов необходим учет всех возможных грамматических образцов и текстовых вариантов.

Кроме того, следует учитывать характерные для научно-технической прозы конструкции определений новых терминов, такие как *Под графемной конструкцией понимается графическая форма, построенная из базисных, проблемно-ориентированных и/или графических конструкций*. Как показало проведенное нами исследование ряда научных текстов из разных предметных областей, такие конструкции содержат фиксированные лексемы и имеют определенную синтаксическую структуру. Для их автоматического распознавания можно описать их в виде специальных лексико-синтаксических шаблонов, подобных тем, что предложены в [Paice, 1993]. Например, вышеприведенная конструкция схематически может быть описана как **«под» T «понимается» NG** ,
где кавычками выделены фиксированные словоформы; **T** − авторский термин (согласованная именная группа с главным словом в форме творительного падежа); **NG** − описание авторского термина (именная группа с главным словом в именительном падеже).

Кроме определений новых терминов, для научно-технической прозы характерны конструкции, раскрывающие семантические связи между терминами – как словарными, так и авторскими. Эти связи могут быть общелогическими («род-вид», «часть-целое») или специальными [Никитина, 1987], как в конструкции вида *Фиксационная норма характеризует степень нормализации раствора*, которая выражает связь «параметр – носитель параметра» между терминами *фиксационная норма* и *степень нормализации раствора*. Такие конструкции также могут быть описаны в виде лексико-синтаксических шаблонов.

Логико-композиционные особенности научного текста

Научный текст организуется как логически взаимосвязанная последовательность речевых (дискурсивных) действий, соответствующих операциям научного мышления [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 вводятся автором текста и эксплицитно помечаются в нем при помощи общенаучных слов и выражений: *в заключение*, *кроме того*, *в действительности* и т.п. Такие слова и выражения называются также дискурсивными (речевыми) маркерами.

В научном тексте наиболее явными маркерами мыслительных операций служат так называемые ментальные перфомативные высказывания, к примеру: *особо подчеркнем*, *далее мы докажем*, *ниже опишем*, *представим это в виде*). Высказывания включают широкий круг ментальных перфомативных глаголов, таких как *выразим*, *рассмотрим*, *предположим* и т.п. [Рябцева, 1992]. Перформативные высказывания не только помечают, но и квалифицируют соответствующий шаг рассуждения и выстраивают содержание текста в форме рассуждения.

Кроме ментальных перформативов, к дискурсивным словам и выражениям относятся маркеры очередности (*во-первых*, *наконец* и др.); коннекторы – союзы и союзные слова (*однако*, *благодаря тому, что* и др.); слова, характеризующие степень объективности информации (*по-видимому*, *возможно* и др.).

Множество дискурсивных маркеров можно разбить на группы в соответствии с выражаемыми ими мыслительными операциями; перечислим основные группы, приводя примеры соответствующих маркеров:

- Констатация и характеризация (*характеризуя*, *укажем, что*).
- Конкретизация и добавление информации (*в частности*, *в дополнение к*).
- Логические операции и причинно-следственные связи (*по этой причине*, *в силу*, *следовательно*).
- Выделение информации, актуализация внимания (*особо подчеркнем*, *перейдем к*, *прежде всего*).
- Определения и допущения (*будем называть*, *предположим, что*).
- Цитирование, иллюстрация и приведение примеров (*к примеру*, *как пишет автор*).
- Обобщение и резюмирование (*суммируя вышесказанное*, *в общем*).
- Классификация, аналогия и сравнение (*с одной стороны*, *аналогично*).
- Выражение мнения и оценивание (*вряд ли*, *целесообразно считать*, *к сожалению*, *кажется*).

Каждой примененной дискурсивной операции в тексте может соответствовать не одно, а несколько последовательных предложений – дискурсивный сегмент текста. Более того, некоторые операции могут

использоваться как средство реализации других операций: например, характеризация производится при помощи примеров и аналогий. В результате включения одной операции в другую происходит вкладывание соответствующих дискурсивных сегментов текста, и формируется иерархическая его структура.

Переход от одной мысли к другой в научном тексте осуществляется не только при помощи дискурсивных слов, но с использованием так называемых общенаучных переменных [Севбо, 1989] – это абстрактные существительные, называющие аппарат научно-познавательной деятельности: *анализ, гипотеза, проблема, аргумент, следствие, идея, понятие, модель* и т.п. Эти существительные играют важную роль в структурно-семантическом упорядочении научной информации и часто употребляются в научных текстах с перформативными глаголами, образуя с ними устойчивые глагольно-именные словосочетания (*подвергнуть анализу, проводить аналогию, привести аргумент*) .

К средствам организации научного текста относятся также такие способы его структуризации, как рубрикация, нумерация, абзацное членение, разбиение на разделы и подразделы.

Все рассмотренные средства структуризации текста могут замещать и дополнять друг друга. Например, рубрикация может использоваться в составе ментального перформатива: *Перечислим основные положения: 1)… 2)…*; в тоже время она может быть замещена маркерами очередности (*во-первых, во-вторых*).

Таким образом, для распознавания логико-композиционной структуры научного текста, т.е. всех употребленных в нем дискурсивных операций и средств структуризации, необходима семантическая классификация общенаучных слов и выражений и ее отображение в словаре.

Словарные средства автоматизированной обработки научного текста

Для процедур терминологического и логико-композиционного анализа научного текста разрабатываются следующие словарные компоненты:

- Терминологический словарь проблемной области: для каждого словарного термина указываются его синонимичные варианты, включая аббревиатуры; для многословных терминов отмечается необходимость согласования составляющих их слов.
- Словарь общенаучной речи, включающий как отдельные слова, так и устойчивые словосочетания. Для словосочетаний описываются их синтаксические характеристики (разрывность/неразрывность и др.) и их семантико-синтаксические валентности; для каждой единицы словаря указывается ее семантическая классификационная группа.

- Лексико-синтаксические шаблоны, описывающие характерные конструкции терминоупотреблений (в частности, определений авторских терминов). Шаблоны фиксируют лексемы конструкций и их грамматическую форму, а также задают синтаксические условия заполнения своих пустых мест (слотов).
- Морфологический словарь основ и неизменяемых слов, в который входят все слова из других словарных компонентов. Каждая единица словаря содержит необходимую грамматическую информацию (часть речи и флективный класс слова), а также ссылки на единицы других словарных компонентов, в состав которых входит данное слово.

Отметим, что только первый из перечисленных словарей зависит от конкретной научной области. В настоящий момент терминологический словарь включает термины из области информатики и вычислительной техники, взятые из текстового словаря [Ершов и др., 1991].

Если компьютерные терминологический и морфологический словари по структуре и составу словарных статей традиционны, то другие словарные компоненты новы и нуждаются потому в более подробном описании.

Словарь русской общенаучной лексики [Большакова, 2002] охватывает широкий круг семантически и грамматически разнородных слов и выражений общенаучной речи, используемых как дискурсивные маркеры. К числу таких выражений относятся именные и глагольно-именные словосочетания с общенаучными переменными (*сильный аргумент*, *привести аргумент*, *опровергнуть гипотезу*), предложно-именные сочетания (*в общих чертах*), причастные и деепричастные обороты (*упомянутый выше*, *резюмируя вышесказанное*), составные предлоги и союзы (*в случае*, *на основе*, *несмотря на то что*).

При построении словаря была проведена классификация слов и выражений: они были разбиты на смысловые группы согласно их роли в тексте (т.е. применяемой дискурсивной операции), без учета их грамматических характеристик. Каждая группа либо является классом синонимичных выражений, либо включает несколько близких по смыслу подгрупп. В общем случае группа содержит несколько синонимичных выражений разной грамматической природы, к примеру, группа следственной связи включает слова и словосочетания *значит*, *итак*, *таким образом*, *тем самым, как видим* и др. В группах встречаются перформативные выражения различных форм (*мы покажем*; *необходимо заметить*; *резюмируя вышесказанное*; *представляется, что*).

Что касается разработанных лексико-синтаксических шаблонов, то большинство их описывает характерные конструкции определений авторских терминов, состоящие из одного предложения. Этими

шаблонами покрывается примерно 60-70% процентов определений, встречающихся в научных текстах.

Каждый лексико-синтаксический шаблон содержит конкретные словоформы (*будем называть*, *под которым понимается*, *определим* и т.п.) – заметим, что соответствующие лексемы входят в словарь общенаучной речи. Шаблоны содержат также свободные, заполняемые места (слоты), обозначаемые специальными символами. К примеру, шаблон вида **NG_{ACC} [«мы»] «будем называть»T_{INS}**
содержит совместно встречающиеся слова **«мы»** и **«будем называть»**, причем слово **«мы»** может отсутствовать; символ **T_{INS}** (заполняемое место) обозначает авторский термин, который должен быть выражен согласованной (в роде, числе и падеже) именной группой, главное слово которой имеет форму творительного падежа; символ **NG_{ACC}** (заполняемое место) − определение или объяснение авторского термина, выражаемое согласованной именной группой, главное слово которой имеет форму винительного падежа, причем эта группа может быть расширена придаточным предложением. Указанный шаблон описывает, например, такое определение авторского термина: *Ресурс, для которого требуется организация разграничения доступа, будем называть защищаемым ресурсом*.

Процедуры терминологического и логико-композиционного анализа текста

Распознавание авторских терминов в тексте и выявление его логико-композиционной структуры включает следующие шаги, каждый из которых реализуется соответствующей процедурой, использующей нужный словарный компонент:

1. Выявление композиционных элементов текста – рубрик, нумерации, абзацев, разделов и подразделов.
2. Распознавание словарных терминов и их комбинаций, при этом выделяются словарные термины максимальной длины (например, термин *системой управления базами данных* будет выделен полностью, хотя в его составе есть другой словарный термин *база данных*).
3. Выделение слов и словосочетаний общенаучной лексики. Например, в предложении *Таким образом, актуальной является задача разработки достаточно полного словаря* будут распознаны общенаучные словосочетания: *таким образом*, *являться актуальной задачей*, *достаточно полный*.
4. Наложение лексико-семантических шаблонов определений терминов и извлечение авторских терминов и определяющих их языковых выражений. Если в обрабатываемом предложении

встретилось слово, входящее в состав некоторого шаблона, то происходит сопоставление предложения с этим шаблоном, при этом проверяются синтаксические условия для его заполняемых мест. В случае успешного сопоставления происходит извлечение языковых конструкций из заполненных мест шаблона. К примеру, в результате успешного сопоставления вышеописанного шаблона (со словоформой **«будем называть»**) с текстом-примером определения будет выделен авторский термин *защищаемый ресурс* и определяющая его конструкция.

5. Выделение именных словосочетаний, которые соответствуют определенным грамматическим образцам терминов – эти словосочетания рассматриваются как кандидаты в авторские термины, введенные автором в текст без определения. Среди выделенных кандидатов определяются возможные синонимичные варианты одного термина.
6. Дискурсивная характеризация предложений, т.е. отнесение каждого предложения текста к определенной дискурсивной операции, базируясь на выделенных в нем на шаге 3 общенаучных выражениях.
7. Определение групп предложений, относящихся к одному и тому же дискурсивному приему. При этом используются эвристики поиска конца сегмента (различные для разных дискурсивных приемов), использующие установленные на шаге 6 дискурсивные характеристики предложений текста и выявленные на шаге 1 композиционные элементы текста.

Описанные процедуры используют поверхностный синтаксический анализ, содержащийся в основном в проверке согласованности слов предложений (например, согласование составляющих слов в употребленных словарных выражениях).

Отметим, что распознанные на шагах 2 и 3 словарные единицы «свертываются», и на последующих шагах анализа не рассматриваются – таким образом сужается область поиска на шаге 5 кандидатов в авторские термины. Решение о том, считать ли некоторую группу синонимичных вариантов одним авторским термином, принимается эвристически.

Результатом последней процедуры является логико-композиционная схема текста, фиксирующая его иерархическую, древесную структуру. Листья древесной структуры соответствуют предложениям текста; нелистовые узлы – разделам/подразделам текста, рубрикам и дискурсивным сегментам; а ветви дерева – структурно-смысловым связям сегментов и предложений (логическим связям и связям подчинения/вхождения).

Заключение

Описанные процедуры и словари составляют основу разрабатываемого в настоящий момент комплекса программ автоматизированного анализа научно-технических текстов, который может быть полезен не только для структурирования и извлечения текстовых знаний, но и для решения других прикладных задач автоматизированной обработки научно-технических текстов – литературно-научного редактирования, реферирования и аннотирования.

Список литературы

[Большакова, 2002] 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.

[Ершов, 1991]Ершов А.П., Шанский Н.М., Окунева А.П., Баско Н.В. Терминологический словарь по основам информатики и вычислительной техники. – М.: Просвещение, 1991.

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

[Никитина, 1987] Никитина С.Е. Семантический анализ языка науки. На материале лингвистики. – М.: Наука, 1987.

[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.

[Осипов, 1997] Осипов Г.С. Приобретение знаний интеллектуальными системами. Основы теории и технологии. – М.: Наука, 1997.

[Пшеничная и др., 1991] Пшеничная Л.Э., Коренга О.Н. Научный термин в словаре и тексте // НТИ. Сер.2. 1991, №12.

[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.

[Bolshakova, 2001] Bolshakova E. Recognition of Author’s Scientific and Technical Terms. In: Computational Linguistics and Intelligent Text Processing. A. Gelbukh (Ed.). Lecture Notes in Computer Science, N 2004, Springer-Verlag, 2001.

[Paice, 1993] 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.