УДК 811.111

В.В. Серебрякова

СТАТИЧЕСКИЙ АСПЕКТ СОДЕРЖАНИЯ ТЕКСТА

Статический аспект основного содержания текста, рассматриваемый в данной статье, лежит в основе такой проблемы, решаемой прикладной лингвистикой, как автоматическая обработка текстов. Рассмотрение лексико-семантической организации текстов по легкой промышленности в сфере «текстиль» происходило на основе статистического метода. Выделенные с целью формального представления смысла текста ключевые слова в дальнейшем позволяют изучать логико-семантическую организацию текстов данной тематики. Статья вносит вклад в решение одной из таких задач лингвистических информационных ресурсов, как ликвидация лакун в области корпусной лингвистики.

Одной из важнейших проблем прикладной лингвистики на сегодняшний день является автоматическая обработка текстов. Благодаря увеличению международных контактов в сфере промышленности и непрерывному обмену информацией, в том числе на английском языке, специалисты все чаще сталкиваются с проблемой невозможности обработки огромного количества документации в срок. К настоящему времени большинство таких программ реализовано в рамках систем машинного перевода. Наличие лексических лакун существенно замедляет процесс обработки информации, к примеру, в такой области, как легкая промышленность, которая является наиболее развитой в нашей республике. В свою очередь, несвоевременно или некорректно переведенные тексты препятствуют продвижению на рынке новейших технологий и разработок в области материаловедения, замедляют процесс обмена опытом в этой сфере.

Анализ исследованной на сегодняшний день литературы показывает отсутствие существующих русско-английских «бумажных» словарей по тематике «Текстиль». Не находит оправдания и тот факт, что при большом числе появляющихся сегодня электронных приложений к словарям и программам машинного перевода электронная база лексики по текстильной промышленности не обновлялась с 1960 года. Разработчики программ упорно продолжают обходить эту тематику стороной, уделяя внимание более приоритетной, такой, как экономическая, юридическая, электронно-вычислительная и лексика нефтяной и газовой промышленности.

Для создания полноценной базы любого электронно-справочного фонда прежде всего необходимо проанализировать весь корпус текстов по той или иной тематике. В рамках такого направления в работе лингвистических информационных ресурсов решение этой проблемы сводится к выделению универсальных категорий текста, что имеет непосредственное отношение к вопросу о его динамике и статике.

Статическое состояние соответствует тексту, рассматриваемому как некоторый результат речемыслительной деятельности, динамическое состояние – тексту в процессе порождения и восприятия [7].

В центр системы языка традиционно ставится слово, и это справедливо при рассмотрении языка с позиции статической системы.

В каждом тексте есть свой словарь и свой синтаксис [4, с. 123; 5, с. 141]. Иногда эти понятия трактуются по-разному. Нам ближе всего точка зрения, где под словарем понимают так называемое слово «содержание» или «несомые слова», а под синтаксисом имеется в виду грамматический строй вместе с «несущими» или служебными сло-

---

Научный руководитель А.В. Зубов – доктор филологических наук, профессор, заведующий кафедрой информатики и прикладной лингвистики УО «Минский государственный лингвистический университет»

вами [9, с. 56]. Таким образом, полагая, что слова «содержание» представляют статику текста и являются отражением в тексте некоторого множества предметов, явлений, фактов реальной действительности, можно рассматривать синтаксис как динамику текста, отражающую те отношения между предметами, фактами, явлениями, которые устанавливает автор текста в зависимости от целевой установки, типа текста, речевого опыта и целого ряда других факторов.

Что же представляет собой статический аспект содержания текста с точки зрения его организации как единого целого?

Понятие «содержание текста» трактуется различными учеными по-разному. К примеру, выделяются содержательно-фабульная, содержательно-концептуальная и содержательно-подтекстовая составляющие [2, с. 27], логическое, предметное (тематическое) и информационное (общесмысловое) содержание. Принимая этот подход, отнесем логическое содержание текста к его динамическому аспекту и остановимся на анализе тематического и информационного содержания текста.

«Основная особенность текста, отличающая его от бессвязного набора фраз, – повторение одинаковых или семантически близких понятий в абзаце» [10, с. 26]. В любом связном тексте мы обнаруживаем многочисленные и разнообразные виды повторяющихся семантических элементов, навязываемых тексту логикой его развития» [1, с. 2].

На сегодняшний день существуют достаточно любопытные попытки формального представления смысла текста через опорные слова и словосочетания [6; 11]. В структуре «опорных слов» или «смысловых вех» существует своя иерархия [12, с. 123]. Главный предмет сообщения выражается в тексте «главными», «ключевыми словами», которые еще иначе называют «номенклатурными дескрипторами» или «выступающими точками» всего содержания [12, с. 123; 7, с. 147]. Эти слова определяют главный предмет сообщения и являются свернутым замыслом текста. При этом один референт может выражаться различными словами – контекстуальными синонимами в пределах приравнивания [12, с. 138]. Из-за линейности текста такие слова могут быть представлены в различных абзацах.

Авторы работ по выделению опорных слов предлагают два формальных критерия определения главных опорных слов (ГОС): их частое употребление (принимая во внимание все контекстуальные синонимы и местоименные замены) и максимальное число абзацев, в которых они встречаются. Далее эти слова становятся ядром, вокруг которого формируются другие элементы, характеризующие микроситуацию (глаголы, прилагательные, наречия и проч.) [8, с. 53].

Главные референты в тексте, однако, чаще всего связаны с другими референтами, которые также незаменимы для описания той же ситуации, но не всей целиком, а, скорее, ее части (микроситуации). Такие слова называют второстепенными опорными словами (ВОС). Встречаемость их по сравнению с ГОС невелика, и число абзацев, содержащих их в тексте, тоже соответственно небольшое.

Проследим процедуру выявления основного статического содержания текстов одной тематики с помощью статистического метода на примере текстов по материаловедению легкой промышленности.

После ввода текста в компьютер был получен распределительный алфавитно-частотный словарь. В словаре содержится информация о частоте употребления каждой словоформы в тексте (F), общем количестве абзацев, в которых встретилась данная словоформа (m), и конкретных номерах абзацев. Все словоформы словаря упорядочены по алфавиту и убыванию частоты их употребления в тексте. Рассмотрим теперь, как выглядит алфавитно-частотный словарь небольшого текста по материаловедению «Nylon» (таблица № 1). В целях экономии мы приводим здесь слова только одной буквы алфавита.

Таблица 1 – Алфавитно-частотный словарь текста “Nylon”.

	F	m	Номера абзацев и частота встречаемости в них слова					
A	4	3	0:2			3:1		5:1
ABILITY	1	1					4:1	
ABOUT	2	2		1:1			4:1	
ABRASION	1	1			2:1			
ACCIDENTAL	1	1					4:1	
ACID	2	2	0:1					5:1
ADDITION	1	1				3:1		
ADIPIC	1	1	0:1					
AFFORDABLE	1	1	0:1					
AFTER	1	1					4:1	
ALSO	1	1						5:1
AMERICAN	1	1	0:1					
AND	11	6	0:2	1:1	2:2	3:1	4:2	5:3
ANY	1	1						5:1
APPLY	1	1						5:1
APPRECIABLY	1	1				3:1		
ARE	4	2			2:3			5:1
ARRESTING	1	1				3:1		
AS	8	3	0:2		2:2			5:4
ASSOCIATED	1	1						5:1
AT	3	2	0:1	1:2				

Затем было сделано следующее:

1. Из полученного словаря была удалена некоторая служебная и общеупотребительная лексика, а именно: были исключены все артикли (a, an, the), предлоги (например; at, for, in, on, to и прочие), союзы (and, but и другие), личные указательные, вопросительные и относительные местоимения (his, that, which), вспомогательные (do, have, will) и модальные (must, should) глаголы, порядковые и количественные числительные (one, two, first, second).

2. Далее различные грамматические формы одного и того же слова были подвергнуты анализу и объединены в группы следующим образом. Например, словоформа «process» другого текста «Wool» имеет частоту F = 3 и встречается в двух абзацах текста. Словоформа «processed» с частотой F = 1 встретилось в одном абзаце текста. Словоформа «processing» имеет свою частоту F = 4 и встречаемость в трех абзацах этого же текста. Из анализа статьи видно, что форма «processing» встречается в функции прилагательного с частотностью F = 3 и существительного с частотностью F = 1 (аналогичным же образом в других текстах той же тематики нам встретилась еще одна форма «processes», которая может обозначать как множественное число существительного “a process”, так и являться формой третьего лица единственного числа Present Simple глагола «to process»). Следовательно, приходится разграничивать эти словоформы. После вычитывания текста и объединения грамматических форм в словаре по принципу принадлежности к частям речи из словоформ «process», «processing» и «processed» вычленили вначале словоформы существительного, а затем глагола и прилагательного. Используя арифметические вычисления, приходим к следующему выводу: из словоформы «process» существительное process(n) встречается в тексте 4 раза в трех абзацах, прилагательное process(a) 3 раза в одном абзаце, а глагол process(v) 1 раз в одном.

3. Следующим этапом стало объединение словарных и контекстуальных синонимов одного и того же слова. Например, словоформа «nylon» имеет частоту F = 13 и встречается в пяти абзацах текста, а словоформа it, которая является контекстуальным синонимом словоформы «nylon», имеет частоту F = 4 и встречается в двух абзацах этого же текста 4:3 и 5:1. После объединения в словаре остается словоформа «nylon» с суммарной частотой F = 13 + 4 = 17 и общим количеством абзацев m = 6.

4. Наконец, на последнем этапе из распределительного алфавитно-частотного словаря были удалены словоформы, встретившиеся только в одном абзаце (m = 1) и передающие, следовательно, основное содержание только данного абзаца.

Оставшиеся после перечисленных выше четырех процедур обработки распределительного алфавитно-частотного словаря словоформы составили словарь потенциальных опорных слов текста «Nylon» (таблица №2)

Таблица 2 – Потенциальный словарь опорных слов текста «Nylon»

	F	m	Номера абзацев и частота встречаемости в них слова					
ACID	2	2	0:1					5:1
CALL	2	2	0:1	1:1				
DEVELOPE	2	2	0:1	1:1				
DUPONT	2	1	0:2					
DYE	4	1						5:4
FIBRE	3	2			2:1			5:1
FINE	3	3	0:1		2:1			5:1
LIKE	2	1	0:2					
MANY	3	3			2:1		4:1	5:1
NEW	2	1	0:2					
NYLON	17	6	0:4	1:4	2:1	3:2	4:3	5:3
POLYESTER	2	1						5:2
PRODUCTION	2	2		1:1				5:1
PROPERTIES	2	2	0:1		2:1			
SILK	3	3	0:1		2:1			5:1
SPIN	2	2			2:1			5:1
STEEL	2	2	0:1		2:1			
STOCKINGS	2	1	0:2					
STRENGTH	2	2				3:1		5:1
STRETCH	2	2			2:1	3:1		
STRONG	2	2	0:1		2:2			
SUCCESS	2	1			2:2			
THERMAL	2	2	0:1	1:1				
TOUGH	2	2	0:1		2:1			
USE	3	3	0:1			3:1		5:3
VERY	3	2			2:2		4:1	
WEIGHT	2	1			2:2			
WIDE	2	2		1:1				5:1
WORLD	2	2	0:1	1:1				

Данный словарь также содержит список словоформ с информацией о частоте употребления каждой словоформы в тексте (F) и общем количестве абзацев, в которых встретилась данная словоформа (m).

Затем для каждого слова из словаря потенциальных опорных слов был вычислен коэффициент важности (коэффициент семантической значимости). Вычисления производились по формуле:

$К_{важ} = F \times m/N \times n$, где

F – абсолютная частота слова в тексте;

m – общее число абзацев, в которых встретилось слово;

N – общее число слов в тексте;

n – общее число абзацев в тексте.

Используя приведенные в работе [2, с. 18–19] формулы, для текста «Nylon» с помощью компьютера были определены критические (пороговые) значения коэффициентов важности слова $К_{важ}^{1}$ и $К_{важ}^{2}$. Они и позволили разделить словарь потенциальных опорных слов текста на две части и получить список главных опорных слов текста (ГОС) и второстепенных опорных слов текста (ВОС). (таблица 3. 4.) Так, для текста «Nylon» главными опорными словами оказались:

Таблица 3 –Главные опорные слова текста «Nylon»

	F	m	Номера абзацев и частота встречаемости в них слова
FINE	3	3	0:1 2:1 5:1
MOST	3	3	2:1 4:1 5:1
NYLON	17	6	0:4 1:4 2:1 3:2 4:3 5:3
SILK	3	3	0:1 2:1 5:1
USE	3	3	0:1 3:1 5:3

Таблица 4 – Второстепенные опорные слова текста «Nylon»

	F	m	Номера абзацев и частота встречаемости в них слова
ACID	2	2	0:1 5:1
CALLED	2	2	0:1 1:1
DEVELOPED	2	2	0:1 1:1
FIBRES	3	2	2:1 5:1
PRODUCTION	2	2	1:1 5:1
PROPERTIES	2	2	0:1 2:1
SPUN	2	2	2:1 5:1
STEEL	2	2	0:1 2:1
STRENGTH	2	2	3:1 5:1
STRETCH	2	2	2:1 3:1
STRONG	2	2	0:1 2:2
THERMAL	2	2	0:1 1:1
TOUGH	2	2	0:1 2:1
VERY	3	2	2:2 4:1
WIDE	2	2	1:1 5:1
WORLD	2	2	0:1 1:1

Перечисленные словоформы составляют основное содержание данного текста.

На основании анализа главных и второстепенных опорных слов была составлена таблица, отражающая основное содержание исследуемого текста. Отобранные компью-

тером слова мы разделили на группы: слова-объекты (предметы), характеристика предмета и действие (таблица 5).

Таблица 5 – Основное статическое содержание текста «Nylon»

Тип слова	Предмет	Характеристика предмета	Действие
ГОС043		FINE	
ГОС007		MOST	
ГОС039	NYLON		
ГОС035	SILK		
ГОС006			USE
ВОС026		ACID	
ВОС085			CALLED
ВОС086		DEVELOPED	
ВОС087	FIBRES		
ВОС037	PRODUCTION		
ВОС038	PROPERTIES		
ВОС088		SPUN	
ВОС089		STEEL	
ВОС090	STRENGTH		
ВОС091			STRETCH
ВОС092		STRONG	
ВОС093		THERMAL	
ВОС094		TOUGH	
ВОС079		VERY	
ВОС095		WIDE	
ВОС096	WORLD		

По полученным опорным словам данного текста можно предположить, что речь идет о производстве нейлоновых волокон (NYLON, PRODUCTION, FIBRES, SPUN), описываются их основные характеристики, подчеркивается прочность, особенно при сравнении с шелком (PROPERTIES, STRENGTH, STRONG, FINE, TOUGH, STRETCH, SILK). Речь также ведется о степени распространенности продукции из нейлона (WIDE, WORLD, USE).

Аналогичным образом были получены ГОС и ВОС всех исследуемых текстов по материаловедению. Эти опорные единицы станут основой для дальнейшего изучения логико-семантической организации таких текстов.

СПИСОК ЛИТЕРАТУРЫ

1. Акишина, А.А. Структура целого текста / А.А. Акишина. – Вып. 2. – М., 1979. – 81 с.

2. Гальперин, И.Р. Текст как объект лингвистического исследования / И.Р. Гальперин. – 3-е изд. – М. : Просвещение, 2005. – 144 с.

3. Зубов, А.В. Основы искусственного интеллекта для лингвистов / А.В. Зубов, И.И. Зубова. – М. : Логос, 2007. – 305 с.

4. Кацнельсон, С.Д. Типология языка и речевое мышление / С.Д. Кацнельсон. – Л. : Наука, 1972.

5. Купина, Н.А. Опыт системно-синтаксического анализа семантики связного текста / Н.А. Купина // Семантика и структура предложения. Лексическая и синтаксическая семантика. – Уфа : Изд-во : БГУ, 1978.

6. Лекомцев, Ю.К. Психическая ситуация, предложение и семантический признак / Ю.К. Лекомцев // Труды по знаковым системам. – VI. Учен. зап. Тартуского ун-та, 1973. – Вып. 308. – С. 5 – 44.

7. Новиков, А.И. Семантика текста и ее формализация / А.И. Новиков. – М. : Наука, 1983. – 216 с.

8. Новиков, А.И. К вопросу о теме и денотате / А.И. Новиков, Г.Д. Чистякова // Известия АН СССР. Сер. Лит. и язык. – 1981. – Т. 40, № 1. – С. 48–56.

9. Проблемы текстуальной лингвистики. – Киев : Изд-во КГУ, 1983.

10. Севбо, И.П. Структура связного текста и автоматизация реферирования / И.П. Севбо. – М., 1969.

11. Скороходько, Э.Ф. Семантические сети и автоматическая обработка текста / Э.Ф. Скороходько. – Київ : Навукова думка, 1983. – 218 с.

12. Лингвистические вопросы алгоритмической обработки сообщений. – М. : Наука, 1983.

Serebryakova V.V. **Static Aspect of the Main Content of the Text**

The static aspect of the main content of the text discussed in this article lies at the heart of such problem solved in the applied linguistics as an automatic text processing. The research of the lexical-semantic organization of the texts on light industry in the area of «textile» was made on the basis of statistical method. The key words marked to visualize the meaning of the text will further allow exploring the logical and semantic organization of the texts on this theme. The article contributes to the solution of one of the tasks of linguistic information resources as the elimination of gaps in the Corpus Linguistics.

Матэрыял паступіў у рэдкалегію 03.07.2009