

ОПРЕДЕЛЕНИЕ ТЕМЫ ТЕКСТА РЕЧИ ПОЛИТИЧЕСКОГО ДЕЯТЕЛЯ

Тема текста является одним из его важных структурных компонентов. Существуют разные трактовки данного понятия. Под темой понимают: 1) содержательный компонент коммуникации; 2) объединяющее начало текста; 3) основной тезис текста, подлежащий дальнейшему развёртыванию; 4) смысловое ядро текста, конденсированное и обобщённое содержание текста; 5) основной компонент смыслового начала текста; 6) содержательный компонент ситуации [1, с. 60]. Понятие «тема» можно рассматривать как в широком, так и в узком смысле. Широкое понимание темы включает понятие сообщаемой в тексте информации, а узкое — название объекта, о котором идёт речь. Но в обоих случаях общей объективной чертой темы является предмет изложения [1, с. 62–63]. Следовательно, тему текста можно представить как отражение в сознании человека наиболее существенных составляющих описанной в тексте ситуации. Это тот инвариант, который остаётся одним и тем же, как для автора текста, так и для его читателя. Т.А. ван Дейк и В. Кинч [2] рассматривают тематическую структуру текста как иерархическую структуру, поскольку тема всего текста может быть описана с помощью более мелких подтем, которые в свою очередь членятся на еще более мелкие микротемы. Каждое предложение текста связано с той или иной подтемой основной темы текста.

Тема устанавливает совокупность определённых слов и словосочетаний, отражающих наиболее существенные составляющие описываемой в тексте ситуации и наиболее важные отношения между ними. Если несколько изменить формулировку Е. Агриколы [3, с. 13], под темой текста будем понимать его понятийное ядро, передающее содержание текста в сжатом и абстрактном виде, в форме смыслового комплекса, выраженного словесными средствами и состоящего из имен героев (или предметов) и предикатов. Слова, выражающие наиболее важные для текста понятия, как правило, употребляются в тексте неоднократно. Они несут основную семантическую нагрузку, обеспечивают становление и развитие темы, связывают отдельные текстовые структуры, создавая тем самым целостность текста. Поэтому их называют ключевыми или опорными словами. Для выбора из текста ключевых слов чаще всего используются статистические, позиционные и лингвосемантические методы [4].

Исходя из цели данного исследования, рассмотрим более подробно статистические методы. Они основаны на использовании статистических параметров для оценки информативности различных элементов текста и учитывают, прежде всего, показатель частоты встречаемости слов в тексте. В результате ранжирования лексики того или иного документа определяются слова с высоким рангом и их сочетаемость в различных фразах [5, с. 144]. Именно статистические методы используются в наши дни в промыш-

ленных компьютерных системах для решения таких задач, как аннотирование и реферирование текстов, их тематическая классификация и кластеризация, порождение и понимание, смысловой поиск и т.д., которые можно отнести к проблеме тематического анализа. Статистическая информация об отдельных лексических единицах легко извлекается компьютером из текста, и есть все основания полагать, что она адекватно отражает его тему (основное содержание).

В настоящее время существует целая группа статистических методов. Так, в одном из них [6, с. 125] процедура выделения ключевых слов текста сводится к следующему. Сначала компьютер разбивает текст на абзацы и составляет по каждому абзацу алфавитно-частотный словарь словоформ. Далее алфавитно-частотные словари абзацев объединяются в распределительный алфавитно-частотный словарь словоформ всего текста. В нем указывается общая частота употребления словоформы в тексте и число абзацев, в которых она встретилась. Затем производится сокращение распределительного алфавитно-частотного словаря словоформ текста до словаря потенциальных ключевых (опорных) слов. Эта процедура включает в себя следующие операции. Сначала из распределительного словаря удаляется вся служебная и общеупотребительная лексика. Затем в оставшейся части словаря суммируются частоты всех грамматических форм одного и того же слова. Далее в этом словаре суммируются частоты синонимов (в том числе и контекстуальных). И затем из оставшейся части распределительного словаря удаляются слова, которые встретились только в одном абзаце. Наконец, на последнем этапе словарь потенциальных опорных слов текста делится на две части. В первую часть входят главные опорные слова (ГОС), а во вторую — второстепенные опорные слова (ВОС). Данная процедура осуществляется с учетом коэффициента важности слова, который вычисляется по формуле:

$$K_v = F * m / N * n,$$

где **F** — абсолютная частота употребления слова; **m** — количество абзацев, в которых встретилось слово; **N** — общее число словоупотреблений в тексте; **n** — общее число абзацев в тексте. Предварительно, в зависимости от длины текста (в словах и абзацах) по специальной формуле определяется средняя (пороговая) величина коэффициента важности [7, с. 45]. Коэффициент важности каждого слова словаря сравнивается с этой величиной. Если коэффициент важности слова выше пороговой величины, то оно относится к числу главных опорных слов, если меньше пороговой величины — к числу второстепенных опорных слов текста. Слова, входящие в каждую из выделенных групп, не однородны по своему содержанию. Они обозначают основные компоненты описанной в тексте ситуации: адресата и адресанта (субъекты), совершаемые ими действия в определенное время и в определенном месте, а также объекты действительности, о которых идет речь в тексте.

Для адекватного формализованного представления темы (основного содержания) необходимо оперировать не отдельными словами, а опреде-

ленным набором ключевых слов, который целесообразно представить в виде таблицы основного статического содержания текста (ТОСС). На основе описанного выше статистического метода были получены ТОСС для десяти инаугурационных речей и девяти поздравительных рождественских речей американских президентов. В качестве примера в табл.1 приводится основное статическое содержание текста инаугурационной речи Б. Клинтона.

Таблица 1

Таблица основного статического содержания текста инаугурационной речи американского президента Б. Клинтона

№ п/п	Тип опорного слова	Опорные слова текста				
		субъекты	объекты	действия	время	место
1.	ГОС 1		<i>world</i>			
2.	ГОС 2			<i>must</i>		
3.	ГОС 3			<i>have</i>		
4.	ГОС 4					<i>America</i>
5.	ГОС 5	<i>people</i>				
6.	ГОС 6			<i>change</i>		
7.	ГОС 7					<i>US</i>
8.	ГОС 8				<i>time</i>	
9.	ГОС 9			<i>let</i>		
10.	ГОС 10			<i>do</i>		
11.	ГОС 11	<i>Americans</i>				
12.	ГОС 12			<i>can</i>		
13.	ГОС 13	<i>fellow</i>				
14.	ВОС 1			<i>will</i>		
15.	ВОС 2			<i>work</i>		
16.	ВОС 3	<i>government</i>				
17.	ВОС 4		<i>generation</i>			
18.	ВОС 5		<i>renewal</i>			
19.	ВОС 6		<i>service</i>			
20.	ВОС 7				<i>season</i>	
21.	ВОС 8			<i>renew</i>		
22.	ВОС 9	<i>millions</i>				
23.	ВОС 10				<i>make</i>	
24.	ВОС 11		<i>freedom</i>			
25.	ВОС 12		<i>way</i>			

Таблицы основного статического содержания всех исследованных текстов вошли в лингвистическую базу знаний системы автоматического порождения текста речи политического деятеля.

ЛИТЕРАТУРА

1. Вишнякова, С.А. Теоретические основы обучения моделированию научного текста (Русский как иностранный, основной этап обучения). — СПб., 2001.
2. Дейк, Т.А. ван. Стратегии понимания связного текста // Новое в зарубежной лингвистике. — М., 1998. — Вып. 23. — С. 153–211.

3. Agricola, E. Vom Text zum Thema. — Berlin, 1976.
4. Горняк, Л.В., Дорохина, Г.В. Тематический фильтр текстов // Искусственный интеллект. — Донецк, 2004. — № 4. — С. 580–586.
5. Блюменау, Д.И. Информационный анализ и синтез для формирования вторичных документов: учебно-практич. пособие. — СПб., 2002.
6. Зубов, А.В. Основы лингвистической информатики: Компьютерная лингвистика: в 3-х ч. — Минск, 1992. — Ч. 2.
7. Зубов, А.В. Вероятностно-алгоритмическая модель порождения текста // Проблемы порождения текста: в 4-х ч. — Минск, 1989. — Ч. 2.