А. С. Семак

СТАТИЧЕСКИЙ АСПЕКТ СОДЕРЖАНИЯ ТЕКСТОВ ОЦЕНОЧНЫХ ОТЗЫВОВ НА ПРЕДПРИЯТИЯ ГОСТИНИЧНОГО ТИПА

Большинство лингвистов отмечает, что текст является многоаспектным, многоплановым явлением. К одному из аспектов текста относят «статику» и «динамику». Внешняя форма текста существует в материальном виде как последовательность лексических единиц. В процессе их декодирования они начинают приобретать содержание, которое формируется в сознании в виде мыслительного образования. Таким образом, текст, как результат речемыслительной деятельности, находится в статическом состоянии. Текст в процессе порождения, восприятия и понимания находится в динамическом состоянии [1, с. 31].

Как отмечают исследователи, в процессе восприятия текста сначала воспринимается семантика предложений, а потом их синтаксическая структура [2]. На семантическом уровне текст состоит из высказываний и сверхфра-

зовых единств. На композиционно-семантическом уровне выделяют главы, параграфы, разделы, абзацы и т.д. [3]. Данные элементы являются взаимозависимыми и образуют смысловое единство всего текста. Как показывают лингвистические исследования смысловое единство текста заключается в определенном наборе «ключевых слов» («смысловых вех», «топиков», «ядер» и т.д.) [4]. Очевидно, что исходя из наличия в тексте таких свойств как «целостность» и «связность» можно утверждать, что в каждом тексте присутствуют конкретные объекты и субъекты в определенном, фиксированном отрезке времени и ограниченном количестве мест. Эти компоненты, объективируемые в «ключевых словах», распределены по тексту неравномерно из-за индивидуальной цели автора текста. Это неравномерное распределение формирует иерархию «ключевых слов», которая соотносится с иерархией денотатов [5, с. 79].

Таким образом, для определения смыслового содержания текста требуется определить его «ключевые слова». Существует три группы методов извлечения ключевых слов, словосочетаний или предложений из текста: статистические, позиционные и логико-семантические. Основной критерий статистических методов – частота употребления знаменательных слов. Позиционные методы исходят из предпосылки, что наиболее значимая информация содержится в заглавии, начале и конце текста. Логико-семантические методы используют целый ряд критериев для извлечения наиболее важной информации. Например, смысловой вес слов, наличие семантической связи между предложениями, употребление в тексте слов из названия, наличие глагола в предложении и т.д. Несмотря на количество методов и критериев отбора ключевых компонентов текста, можно отметить, что они ориентированы непосредственно на результат этого процесса, а не на моделирование самого процесса понимания текста человеком.

Несмотря на недостатки вышеуказанных методов, стоит заметить, что они обладают рядом достоинств. Статистический метод отличается высокой точностью результатов, поэтому именно он послужил методом определения «ключевых слов» в нашем исследовании. В качестве материала исследования послужили оценочные отзывы на предприятия гостиничного типа. Под «оценочным отзывом» понимается оформленный тип текста, содержащий авторскую оценку о некоем явлении действительности, размещенный в сети Интернет (или на других ресурсах) с целью обмена информацией между пользователями. Причиной выбора оценочных отзывов для анализа стала активизация лингвистических исследований в области определения тональности. Системы определения тональности широко применяются в сфере мониторинга качества услуг и товаров, бизнес-аналитики, кинопроката и политического прогнозирования. Под определением тональности понимается определение характера субъективной информации, содержащейся в тексте. Тональность может быть положительной, нейтральной и отрицательной. На настоящий момент количество методов определения тональности очень велико, но их принято распределять на четыре группы: методы, основанные на правилах, основанные на словарях, основанные на машинном обучении с учителем и на машинном обучении без учителя. Каждая группа методов

обладает рядом достоинств и недостатков, характерных для различных задач анализа тональности. Стоит заметить, что ни один из вышеуказанных методов не рассматривает оценочный отзыв как связную структуру. Например, в методе, основанном на словарях тональной лексики, происходит сравнение определенных лексических единиц с единицами тонального словаря: если прилагательное, встречающееся в тексте, имеет в таком словаре положительную тональность, то тональность существительного, к которому относится это прилагательное, так же будет положительной. Методы, основанные на машинном обучении, исходят из принципа использования заранее размеченных текстов-образцов, на основе которых происходит обучение компьютерного классификатора. Очевидно, что при таких подходах наличие смысловых связей внутри текста игнорируется во время анализа. Таким образом, для определения тональности посредством формализации семантики текста требуется определить его статику и динамику.

Для последующего анализа нами были использованы отзывы с оценочных Интернет сайтов booking.com, tripadvisor.com Причиной выбора этих сайтов послужила политика администрации этих ресурсов направленная на предотвращение публикации информации коммерческого содержания. Например, сайт www.booking.com позволяет оставлять отзыв об отеле только после фактической оплаты пребывания в нем.

Целью статистического метода являлось определение основных объектов текста: действующие лица, действия, состояния, места, время и т.д. Статистический анализ показал, что на протяжении всего текста употребляются опорные слова, которые раскрывают тему. Распределение данных слов не является равнозначным. Слова с наибольшим семантическим весом являются *главными опорными словами* (ГОС). Слова, раскрывающие отдельные микроситуации являются *второстепенными опорными словами* (ВОС). Статическое содержание всего текста отражено в последовательности ГОС и ВОС [5].

В качестве основной логико-семантической единицы текста в нашем случае выступает абзац, т.к. именно в предметно-логическом содержании абзаца текста раскрывается его микротема. Под предметно-логическим содержанием абзаца понимается перечисление фактов, объектов действительности так или иначе связанных с сообщением текста. В независимости от жанра текста, абзац строится по определенным правилам, которые отражают структурно-смысловые модели изложения мыслей [6, с. 335]. Данные модели усваиваются человеком еще в раннем детстве и хранятся в памяти в виде языковых шаблонов [7]. Они представляют собой элементы психической реальности, которые напрямую участвуют в порождении высказывания. Причем, синтаксические шаблоны могут охватывать больше одного предложения и формировать шаблон-высказывание. Количество и специфичность таких шаблонов определяется текстами одной предметной области или одного автора. Совокупность таких шаблонов-высказываний в лингвистике называется «сложным синтаксическим целым» или «сверхфразовым единством». Но из-за отсутствия формальных признаков таких составляющих подвергнуть компьютерной обработке сверхфразовое единство и сложное синтаксическое целое чрезвычайно трудно.

Исходя из вышесказанного, можно сделать вывод, что абзац является отражением фрагмента психической ситуации и формирует смысловое и логическое единство текста, что подтверждается рядом психолингвистических экспериментов [8, 9]. В ходе данных экспериментов было установлено, что членение текстов на абзацы происходит не автоматически, а носит целенаправленный характер. Так же абзац исключительно удобен для формализации из-за наличия формального признака (красной строки). Таким образом, текст можно представить в виде последовательного изложения его абзацев.

На начальном этапе анализа отзывов был составлен алфавитно-частотный словарь каждого текста с указанием общего количества словоупотреблений и количество абзацев с данным словом. Далее алфавитно-частотного словаря были удалены все служебные и общеупотребительные слова. Например, были удалены предлоги, союзы, артикли, вспомогательные и модальные глаголы и числительные. Показатели были скорректированы с учетом объединения словоформ, лексических повторов и различных видов синонимий. На последнем этапе обработки алфавитно-частотного словаря были удалены все словоформы, употребленные только в одном абзаце (m = 1), т.к. они относятся к содержанию только одного абзаца, а не всего текста. В итоге был получен словарь потенциальных опорных слов. Дальнейшее распределение на ГОС и ВОС осуществлялось по формуле:

$$К_{важ} = \frac{(F*m)}{(N*n)},$$

где:

F – абсолютная частота слова в тексте;

m – число абзацев, в которых встретилось слово;

N – общее число слов в тексте;

n – общее число абзацев в тексте.

Извлечение ГОС и ВОС происходит на основе сравнения коэффициента важности с граничными значениями $К_{1важ}$ и $К_{2важ}$, которые вычисляются для каждого текста исходя из количества абзацев и общего числа словоформ.

Например, в одном[1] из анализируемых оценочных отзывах имеется 6 абзацев и 359 слов. В ходе вышеописанного анализа было выделено 4 ГОС (табл. 1) и 10 ВОС (табл. 2).

Таблица 1

Главные опорные слова оценочного отзыва

№ слова	ГОС	Общее количество словоупотреблений	Общее количество абзацев, где встретилось слово
1	We	13	4
2	Breakfast	7	4
3	Room	6	4
4	Hotel	5	4

[1] Отзыв размещен на сайте: http://www.tripadvisor.com

Таблица 2

Второстепенные опорные слова оценочного отзыва

№ слова	ВОС	Общее количество словоупотреблений	Общее количество абзацев, где встретилось слово
1	Night	3	2
2	Pounds	3	2
3	Check-in	2	2
4	Could	2	2
5	Got	2	2
6	Morning	2	2
7	Took	2	2
8	Usual	2	2
9	Very	2	2
10	Went	2	2

Далее в каждом абзаце текста анализировалась позиция ГОС и ВОС и то, как они раскрывают тему текста. Таким образом, предметно-логическое содержание данного отзыва можно представить в виде цепочки микротем (табл. 3). Каждой микротеме был присвоен код. Всего, исходя из результатов подобного анализа сорока оценочных отзывов, было выявлено 26 типов микротем.

Таблица 3

Предметно-логическое содержание абзацев оценочного отзыва

№ Пункта	Код микротемы	Предметно-логическое содержание абзацев текста
1	М001	Описание процесса бронирования ГОС 3 в ГОС 4
2	М002	Оценочная характеристика ГОС 4
3	М003	Оценочная характеристика ГОС 3 в ГОС 5
4	М004	Общая оценочная характеристика ГОС 3
5	М005	Описание процесса выписки ГОС 1 из ГОС 5
6	М006	Итоговая характеристика ГОС 5 и ГОС 4

Полученные данные отражают статический аспект текстов оценочных отзывов и необходимы для дальнейшей формализации логико-семантических отношений.

ЛИТЕРАТУРА

1. *Новиков, А. И.* Семантика текста и ее формализация / А. И. Новиков. – М. : Наука, 1983. – 215 с.
2. *Леонтьев, А. А.* Психолингвистические единицы и порождение речевого высказывания / А. А. Леонтьев. – М. : Наука, 1969. – 307 с.
3. *Валгина, Н. С.* Теория текста / Н. С. Валгина. – М. : Логос, 2003 – 278 с.
4. *Вейнрих, У.* Опыт семантической теории / У. Вейнрих // Новое в зарубежной лингвистике. – М., 1981. – Вып. 10. – С. 50–176.
5. *Зубов, А. В.* Статистический аспект содержания текста и его формальное выражение / А. В. Зубов // Квантитативная лингвистика и автоматический анализ текста : учен. зап. Тарт. ун-та. – Тарту, 1986. – Вып. 745. – С. 75–91.

6. *Зубов, А. В.* Вероятностно-алгоритмическое моделирование статической и динамической составляющих содержания текста / А. В. Зубов // OSTIS 2011: матер. межд. науч.-технич. конф., Минск, 10–12 февр. 2011 г. / редкол. : В. В. Голенков (отв. ред.) [и др.]. – Минск : БГУИР, 2011. – С. 331–342.
7. *Адмони, В. Г.* Синтаксическая семантика – это семантика синтаксических структур / В. Г. Адмони // Проблемы синтаксической семантики. – М., 1976. – С. 3–8.
8. *Страхова, В. С.* Внешние средства организации текста / В. С. Страхова // Лингвистика текста : сб. науч. тр. МГПИИЯ. – М., 1971. – Вып. 141. – С. 156.
9. *Бондаренко, Г. В.* Текст – смысл – ситуация (к постановке проблемы) / Г. В. Бондаренко, Ю. А. Шрейдер // Вопросы информационной теории и практики – М., 1978. – Вып. 36. – С. 80–91.