АНАЛИТИЧЕСКОЕ АННОТИРОВАНИЕ ТЕКСТОВ СМИ: ЦЕННОСТИ И ОЦЕНКИ

PRODUCING ANALYTIC ABSTRACTS OF MASS-MEDIA TEXTS: VALUES AND ESTIMATES

Зевахина Т.С.
Московский государственный университет им. М.В. Ломоносова, филологический факультет
mailto:tzev@mail.ru,

Олейникова Е.Е.
Телеканал «Россия»
mailto:helen0203@yandex.ru

Описывается опыт разработки и экспериментального опробования системы аналитического аннотирования текстов СМИ в двух взаимосвязанных измерениях – ценностном и оценочном. Результирующий образ текста телевизионного комментария строится с помощью Базового семантического словаря.

1. Корпус текстов Владимира Лусканова

Средства массовой информации (СМИ) точнее было бы называть средствами массового воздействия (СМВ). Коммуникативные явления этого плана исследуются ныне широким фронтом при активном участии лингвистов [1 – 6].

Среди СМВ немаловажную роль играет такой жанр, как телевизионный комментарий к новостям. Эмпирическая база нашего многоэтапного исследования включает, в частности, тексты, написанные или произнесенные с экрана телевизора В.И. Лускановым с 1995 по 2001 год (в дальнейшем Корпус-ВЛ). Автор – признанный авторитет в телевизионной «тусовке», лауреат ТЭФИ и т.д. Мы исходим из гипотезы, что его успех обусловлен именно текстами (а не мастерством телеоператора и монтажера). Мы пытаемся ответить на вопрос – «Как же это у него получается?».

В течение описываемого времени автор работал в различных телекомпаниях (РТР и НТВ), делал телесюжеты, писал для Интеренет-изданий. Наш репрезентативный корпус из 100 текстов включает значительное число репортажей с мест событий, а также сюжетов, выходивших под рубриками «ТЕМА ДНЯ» и «ОСОБОЕ МНЕНИЕ».

Всего корпус содержит 12087 разных слов. С помощью прикладной компьютерной системы мы сводим все слова к гиперлексемам, которые представлены в словаре квазиосновами, и таких квазиоснов насчитывается 4100.

Хронологически и в зависимости от субстанции воспроизведения выделяется четыре группы текстов: НТВ-ЭФИР, НТВ-САЙТ, РТР-ЭФИР, РТР-САЙТ.

Особенности текстов, соотнесенных с видеорядом, подробно описаны в литературе. Сюжеты В.И. Лусканова не являются новостями в узком смысле, однако специфику телевизионных текстов сохраняют (не выраженная в явном виде в тексте информация передается через видеоряд, речь третьих лиц (синхроны) используется для передачи наиболее значимой или спорной информации и т.п.).

Тексты на сайтах – аналог письменной газетной речи. Однако отсутствие не только цензуры любого вида, но и редакторской правки делает их гораздо более «лично авторскими», чем любой текст в традиционных СМИ. Это характерно в большей мере для раннего этапа развития Интернета, к которому относятся анализируемые тексты. Для исследуемых текстов не удалось выявить существенных отличий между эфирными и интернет-текстами. Интеренет-тексты несколько длиннее эфирных, однако, тексты, написанные на РТР, гораздо короче текстов НТВ. Средняя длина текстов - 420 слов (около двух с половиной минут в эфире – стандартное время сюжета в информационных программах.) Самый короткий текст – 86 слов (НТВ-САЙТ 10.03.2000- «Смерть Артема Боровика»). Самый длинный текст – 1194 слова (НТВ-ЭФИР 06.10.1996 – «Конец войны в Чечне» (8 синхронов!)).

Некоторые формальные отличия Интернет-текстов от эфирных аналогов:

- Использование большего количества цифр, чем это было бы допустимо в эфирном варианте:

«*Российские военные в Чечне запутались. Профессия такая. Или недолет или перелет. Математика неточности не допускает. 4 февраля 2000 года начальник Генерального штаба Вооруженных сил РФ генерал-полковник Валерий Манилов сообщил, что с 1 октября в Чечне погибло 882 человека из Министерства обороны и 241 человек из МВД. Итого: 1123. Раненых, по утверждению Манилова, среди военных было 2327, среди сотрудников МВД 854. Итого: 3181.*»
(НТВ-САЙТ 12.02.2000)

- использование традиционного цитирования вместо телевизионных синхронов.

Некоторые дополнительные количественные данные: длина Корпуса-ВЛ 42355 словоупотреблений; повествовательных предложений – 3400; вопросительных – 198; восклицательных – 24; всего 3622; средняя длина предложения – 11, 7 словоупотреблений; количество запятых 3 877.

2. *Компьютерная система извлечения тезаурусных знаний из текста: ценности и оценки*

2.1. Принципы и архитектура

Описываемый здесь подход к прагматико-семантическому анализу текста СМИ является развитием более ранних моделей, предложенных нами для других проблемных областей – для автоматизированной системы управления (директивные тексты), для информационно-поисковой системы (научно-технические тексты) и для экспертной системы (научно-прогностические тексты) [7].

Для многих приложений лингвистической семантики ключевое значение имеет проблема ВЫЯВЛЕНИЯ ЗНАНИЙ И УБЕЖДЕНИЙ, которыми оперируют коммуниканты в процессе общения. Интенсивно разрабатываются такие аспекты этой проблемы, как выбор источников знаний, методы получения знаний из этих источников, классификация типов знаний, оценка их релевантности и достоверности, методы единообразного представления знаний, способы интеграции и обобщения знаний, пути пополнения и коррекции базы знаний.

К числу наиболее сложных и перспективных задач относится задача ИЗВЛЕЧЕНИЯ ЗНАНИЙ ИЗ ТЕКСТА, рассматриваемая как одна из важнейших функций лингвистического процессора, или автоматизированной системы обработки текста (АСОТ) [8]. Применительно к текстам СМВ реализация такой функции позволяет, с одной стороны, моделировать процесс понимания (смыслового восприятия) текстов слушателями и читателями, а с другой стороны, моделировать процесс вербализации коммуникативного замысла журналистом.

Среди многообразных знаний, содержащихся в тексте, важный когнитивный пласт составляют знания, образующие ТЕЗАУРУС ТЕКСТА. Это система текстовых понятий, организованная с учетом (а) релевантных для данной предметной области семантических классов, (б) авторской картины мира и (в) специфики той фактической информации, которую несет именно данный текст. В современном понимании термин “тезаурус” обозначает многоаспектную систему семантических полей лексических (или других языковых) единиц. В излагаемом здесь исследовании моделируются ценностный и оценочный аспекты тезауруса.

КОМПЬЮТЕРНАЯ РЕАЛИЗАЦИЯ извлечения тезаурусных знаний из текста первоначально осуществлялась нами в ходе разработки лингвистического обеспечения экспертной системы на основе принципов, предложенных Б.Ю. Городецким и Г.С. Осиповым. Конкретно-лингвистическая концепция нашей модели сочетает в себе идеи контент-анализа дискурса, тезаурусной систематизации понятий, компонентного анализа лексических значений, аксиологической семантики, прикладного словообразовательного анализа и лексикостатистики.

Созданная типовая компьютерная система включает четыре МОДУЛЯ: (1) ведение опорных словарей; (2) прикладной морфологический анализ словоформ текста (первая версия разработана И.А. Муравьевой); (3) терминологический анализ текста (разработан Б.Ю. Городецким и О.М. Сазоновой); (4) собственно извлечение тезаурусных знаний из текста. Программирование осуществлялось Е.В. Комаровой.

Модуль 4 работает в нескольких РЕЖИМАХ, каждый из которых строит для предложенного текста документа определенный тезаурусный образ. Режимы отличаются друг от друга аспектами и глубиной анализа тезаурусных знаний. Но все они опираются на прикладное тезаурусное моделирование базовой лексики. Основную роль в работе модуля 4 играет Базовый Семантический Словарь (БАСС) – развиваемый словарь важнейших лексических единиц, существенных для данной предметной области. Каждая словарная статья описывает гиперлексему, то есть класс лексем, которые имеют общую квазиоснову и отождествляются в рамках данной понятийной системы. В отдельных зонах словарной статьи описаны аспекты тезаурусной характеристики гиперлексемы. Нами создается вариант БАСС для области телевизионного дискурса (жанр политического комментария). Этот БАСС (наряду с другими, вспомогательными словарями) используется программной системой в ходе автоматического анализа того или иного текста, подаваемого на вход системы. Результатом анализа является Тезаурусный образ текста (ТОТ), формализующий ценностный и оценочный аспекты содержания.

Тезаурусный модуль может рассматриваться как своего рода интерфейс между когнитивным миром авторских текстов и массовым сознанием слушателей (читателей).

При анализе текстов В. Лусканова системный БАСС использует, во-первых, 12 имплицитных обобщенных ценностных категорий, во-вторых, открытое множество эксплицитных ценностных категорий, в-третьих, пометы об отрицательной или

положительной оценке (примеры даны ниже). Пометы первого вида могут быть приписаны в словаре самым разным в сигнификативном и денотативном отношении лексемам, но имплицитно содержащим в себе одну или более обобщенных ценностных категорий (ср. широкий подход к имплицитности в коллективной монографии [Имплицитность 1999]). Эти пометы даются прописными буквами и соединяются (в случае необходимости) знаком "&". Пометы второго вида приписаны (ровно по одной) тем гиперлексемам, которые прямо называют ту или иную (иногда очень конкретную) ценностную категорию. Эти пометы даны прописными буквами в квадратных скобках. Пометы третьего вида приписываются тем лексемам, которые в сигнификативном или экспрессивном слое своего значения содержат весьма интенсивную (понятийную или эмоциональную) оценку - либо отрицательного, либо положительного характера. Эти пометы даны в БАСС строчными буквами в угловых скобках.

Первые два вида помет, или семантических компонентов, используются для построения ценностного ТОТ. Он содержит определенным образом упорядоченные ценностные категории с их статистическим весом. Третий вид помет служит для построения оценочного ТОТ. В нем приводятся конкретные лексемы, снабженные числовым показателем абсолютной частоты. Сущность работы основной программы заключается в идентификации в тексте базовых квазиоснов, в подсчете их частоты и в выписывании информации из БАСС. (При этом программа прикладного морфологического анализа опознает и отбрасывает служебные слова и имена собственные - на данном этапе исследования мы их не подвергаем тезаурусной обработке.) Кроме того, в нашей системе есть программа выделения устойчивых именных словосочетаний, которая дает по каждому тексту дополнительный источник сведений о возможной развернутой номинации ценностей и оценок.

Ниже приводятся образцы обработки двух документов из нашего Корпуса текстов. Текст ВЛ1 звучал 22.08.2000 и был посвящен подводной лодке "Курск", а текст ВЛ51 отражает тему предвыборной президентской кампании (эфир – 11.02.1996).

2.2. Компьютерный анализ документа ВЛ1

Приведем фрагмент составленного системой промежуточного Рабочего алфавитно-частотного ценностно-оценочного словаря квазиоснов для документа ВЛ1:

бед 1 ЖИЗНЬ & СТАБИЛЬНОСТЬ <отр. оц. ситуации>

бог 1 СТАБИЛЬНОСТЬ & ЗАБОТА [БОГ]

бюдж 1 БОГАТСТВО

вер 1 ЗНАНИЕ [ВЕРА]

видн 1 ЗНАНИЕ & СОСТЯЗАТЕЛЬНОСТЬ & СИЛА

вин 1 ОТВЕТСТВЕННОСТЬ <отр. оц. человека>

гибел 2 ЖИЗНЬ <отр. оц. ситуации>

глубоководн 1 ЖИЗНЬ & СИЛА

дет 3 ЖИЗНЬ & ЗАБОТА [ДЕТИ]

давно 1 ЗНАНИЕ

договори 1 СТАБИЛЬНОСТЬ

дом 1 БОГАТСТВО & РОДИНА & ЗАБОТА [ДОМ]

жертвенн 1 ЗАБОТА & ЖИЗНЬ [ЖЕРТВЕННОСТЬ] <полож. оц. человека>

защи 2 ЗАБОТА & ЖИЗНЬ & СИЛА <полож. оц. человека>

зна 1 ЗНАНИЕ [ЗНАНИЕ]

идеал 2 ЗНАНИЕ <полож. оц. ситуации, человека>

истор 1 ЗНАНИЕ & ОТВЕТСТВЕННОСТЬ

капит 2 БОГАТСТВО

мертв 1 ЖИЗНЬ <отр. оц. ситуации>

Приведем статистические данные о соотношении словаря и текста в документе ВЛ1 Прежде всего о составе полученного лексикона квазиоснов.

Число знаменательных квазиоснов - 168.

Число квазиоснов, несущих имплицитную (скрытую) ценностную нагрузку, – 153 (91 %). Как говорилось выше, эта нагрузка сводится к обобщенным ценностным категориям.

Из них число квазиоснов, выражающих, кроме того, и некоторое эксплицитное (прямое) ценностное значение, – 17 (10,1 % от общего числа знаменательных квазиоснов; 11 % от числа ценностно нагруженных квазиоснов). Как мы говорили, в этом случае ценностная категория носит более конкретный характер.

Число квазиоснов, не имеющих ценностной нагрузки, – 15 (8,9 % от общего числа знаменательных квазиоснов).

Что касается абсолютных частот употребления интересующих нас групп лексики, то эти данные выглядят следующим образом.

Длина текста (число словоупотреблений) – 420.

Суммарная абсолютная частота знаменательной лексики в данном тексте – 206 (ее суммарная относительная частота по отношению к общей длине текста - 0, 490, т.е. 49 %).

Суммарная абсолютная частота ценностно нагруженной лексики (несущей имплицитную ценностную нагрузку) – 188 (ее суммарная относительная частота на множестве словоупотреблений знаменательных слов текста 0, 913, т.е. 91,3 %).

Из них частотность 1 имеют 128 квазиоснов, частотность 2 – 17 квазиоснов, частотность 3 – 6 квазиоснов, частотность 4 – 2 квазиосновы.

Суммарная абсолютная частота прямых однословных номинаций конкретного ценностного значения – 23 (суммарная относительная частота на множестве словоупотреблений знаменательных слов текста - 0, 112, т.е. 11,2 %).

Суммарная абсолютная частота знаменательной лексики, не имеющей ценностной нагрузки - 18 (суммарная относительная частота - 0, 087, т.е. 8, 7 %).

Теперь приведем построенный системой итоговый **Ценностный тезаурусный образ документа ВЛ1** (он включает имплицитную и эксплицитную части).

N	имплицитные ценностные категории	абс. част.	Отн. част. к знам. лекс.	отн. част. к ценн. лекс.
1	ЖИЗНЬ	66	0,320	0,351
2	ЗНАНИЕ	53	0,257	0,282
3	ОТВЕТСТ-ВЕННОСТЬ	33	0,160	0,176
4	ЗАБОТА	22	0,107	0,117
5	СИЛА	19	0,092	0,101
6	СТАБИЛЬНОСТЬ	19	0,092	0,101
7	РОДИНА	16	0,078	0,085
8	ВЛАСТЬ	14	0,068	0,074
9	НОВИЗНА	14	0,068	0,074
10	БОГАТСТВО	12	0,058	0,064

Имплицитные ценностные категории ВЛ1

Эксплицитные ценностные категории ВЛ1:

ДЕТИ (4), РОССИЯ (3), РУССКИЕ (2),
БОГ, ВЕРА, ДОМ, ЖЕРТВЕННОСТЬ, ЖИЗНЬ,
ЗНАНИЕ, ЛЮДИ, ОТВЕТСТВЕННОСТЬ,
ПОДВИГ, ПОРЯДОК, РАБОТА,
САМОСТОЯТЕЛЬНОСТЬ, СИЛА, ЧЕСТЬ

Одновременно система строит итоговый ***Оценочный тезаурусный образ документа ВЛ1*** (он включает отрицательную и положительную части).

Отрицательная оценка человека или ситуации реализована следующими гиперлексемами:
гибель (2), *переживания* (2), *разбитый* (2), *бастовать*, *беда*, *вина*, *гибель*, *мертвый*, *могила*, *ошибка*, *перекрывать*, *стучать*, *тяжелый*, *ужас*.

Положительная оценка человека или ситуации реализована с помощью гиперлексем:
спасать (3), *защита* (2), *идеальность* (2), *жертвенность*, *легендарность*, *несравненный*, *переносить*, *подвиг*, *честь*.

2.3. Компьютерный анализ документа ВЛ51

Приведем маленький фрагмент построенного системой промежуточного Рабочего алфавитно-частотного ценностно-оценочного словаря квазиоснов для документа ВЛ51:

банкр 1 БОГАТСТВО & СТАБИЛЬНОСТЬ & ОТВЕТСТВЕННОСТЬ <отр. оц. ситуации, человека>

бед 1 СТАБИЛЬНОСТЬ & ЖИЗНЬ <отр. оц. ситуации>

выплат 1 БОГАТСТВО & ВЛАСТЬ & ОТВЕТСТВЕННОСТЬ & ЗАБОТА

лезт 1 СОСТЯЗАТЕЛЬНОСТЬ & СИЛА & СТАБИЛЬНОСТЬ <отр. оц. человека>

народ 1 РОДИНА & ТРУД & ОТВЕТСТВЕННОСТЬ [НАРОД]

Число квазиоснов, несущих имплицитную (скрытую) ценностную нагрузку, - 187 (92,1%).

Суммарная абсолютная частота ценностно нагруженной лексики – 270 (ее суммарная относительная частота на множестве словоупотреблений знаменательных слов текста - 0,941, т.е. 94,1%).

Приведем построенный системой итоговый **Ценностный тезаурусный образ документа ВЛ51** (его имплицитную и эксплицитную части).

N	Имплицитные ценностные категории	Абс. част.	Отн. част. к знам. лекс.	Отн. част. к ценн. лекс.
1	ЗНАНИЕ	128	0,446	0,474
2	ОТВЕТСТ-ВЕННОСТЬ	84	0,293	0,311
3	СОСТЯЗА-ТЕЛЬНОСТЬ	68	0,237	0,252
4	ВЛАСТЬ	59	0,206	0,219
5	РОДИНА	40	0,139	0,148
6	СТАБИЛЬ-НОСТЬ	40	0,139	0,148
7	НОВИЗНА	32	0,111	0,119
8	БОГАТСТВО	28	0,098	0,104
9	СИЛА	26	0,091	0,096
10	ЗАБОТА	12	0,042	0,044

Имплицитные ценностные категории ВЛ51

Эксплицитные ценностные категории ВЛ51:

БЕЗОПАСНОСТЬ (2), КОМПРОМИСС (2),
ЛИЧНОСТЬ (2), ДЕМОКРАТИЯ (2),
ДОЛГ, ДОМ, ЗНАНИЕ, ИНИЦИАТИВА, НАРОД,
ПРАВДА, РАБОТА, РОССИЯ, СОВЕСТЬ,
СЧАСТЬЕ, УДАЧА.

Оценочный тезаурусный образ документа ВЛ51 (с его отрицательной и положительной частями) выглядит следующим образом.

Отрицательная оценка человека или ситуации реализована следующими гиперлексемами: *пинки* (2), *провал* (2), *проигрыш* (2), *банкротство*, *беда, война, дудаевцы, патовая, популист, пустота, резня, устарелость*.

Положительная оценка человека или ситуации реализована с помощью гиперлексем: *компромисс* (2), *инициативность*, *положительный, помощь, понравиться, правда, сила, совесть, счастье, удача, уникальность*.

3. Интерпретация результатов автоматического тезаурусного анализа текста

Ограниченный объем настоящей статьи не позволяет дать развернутую интерпретацию результатов автоматического ценностно-оценочного тезаурусного анализа предложенных двух текстов. Но постараемся наметить ряд направлений такой интерпретации.

Сопоставление ценностных профилей двух текстов позволяет моделировать как общие установки данного автора, так и различия в коммуникативно-когнитивном заряде этих документов.

Так, в первую половину списка обобщенных имплицитных категорий и в том и в другом профиле попали три следующих ценности: ЗНАНИЕ, ОТВЕТСТВЕННОСТЬ, СТАБИЛЬНОСТЬ. Может быть, один из секретов Лусканова состоит в том, что он делает акцент именно на этих, конструктивных ценностях, привлекая к ним внимание с помощью разнообразных лексических средств (не только прямых, но и косвенных, не только положительных, но и отрицательных)? В этой связи полезно обратить внимание и на категорию НОВИЗНА, которая в семантике обоих текстов занимает среднее (и далеко не последнее место). И это помогает нам увидеть особую черту В. Лусканова как говорящего - его динамизм, наступательность, призыв к энергичному поиску решений.

Дифференцирующие признаки в первой шестерке выглядят так: в тексте ВЛ1 (о трагедии подлодки) внимание привлекается прежде всего к таким категориям, как ЖИЗНЬ, ЗАБОТА, СИЛА, а в тексте о предвыборной кампании – к категориям СОСТЯЗАТЕЛЬНОСТЬ, ВЛАСТЬ, РОДИНА.

Как видим, уже парадигматический анализ лексических средств общения позволяет приоткрыть завесу тайны убедительной аргументации – это выделение стержневых ценностных факторов, это использование разнообразной лексики, группируемой вокруг ценностей, от которых зависит сама ЖИЗНЬ человека и СТАБИЛЬНОСТЬ мира.

Интересное направление открывается при изучении составных ценностей, т.е. сочетаний нескольких категорий, скрытых в глубинах лексических пластов текста. Так, в рабочем ценностном словаре текста ВЛ1 частыми составными ценностями являются ЖИЗНЬ & ЗАБОТА, ЖИЗНЬ & СТАБИЛЬНОСТЬ, ЖИЗНЬ & СИЛА, а в тексте ВЛ51 – ЗНАНИЕ & СТАБИЛЬНОСТЬ, ЗНАНИЕ & СОСТЯЗАТЕЛЬНОСТЬ, ЗНАНИЕ & НОВИЗНА & ОТВЕТСТВЕННОСТЬ.

Весьма рельефное представление о деталях ценностного содержания текстов дают выявленные эксплицитные категории. Так, для текста ВЛ1 выделяются прежде всего ДЕТИ, РОССИЯ, РУССКИЕ и далее – БОГ, ВЕРА, ЖЕРТВЕННОСТЬ, ЧЕСТЬ и др. Для текста ВЛ51 это прежде всего БЕЗОПАСНОСТЬ, КОМПРОМИСС, ЛИЧНОСТЬ, ДЕМОКРАТИЯ и далее – ДОЛГ, ИНИЦИАТИВА, СЧАСТЬЕ, УДАЧА и др.

Оценочные профили каждого текста показывают точность словесных квалификаций, даваемых автором текущим событиям, и могут детально интерпретироваться с помощью дальнейшего компонентного и прагматического анализа, т.е. позволяют от лингвистического моделирования переходить к моделированию политологическому. Лексика здесь говорит сама за себя: текст ВЛ1: *гибель, разбитость, беда, вина, ошибка, ужас; защита, жертвенность, честь*; текст ВЛ51: *пинки, провал, проигрыш, банкротство, устарелость; компромисс, правда, совесть*.

Литература

1. Зимбардо Ф., Ляйппе М. Социальное влияние. – С.-Петербург, 2001.
2. Имплицитность в языке и речи / Отв. ред. Борисова Е.Г., Мартемьянов Ю.С. – М., 1999.
3. Почепцов Г.Г. Теория и практика коммуникации (От речей президентов до переговоров с террористами). – М., 1998.
4. Язык СМИ как объект междисциплинарного исследования / Отв. ред. Володина М.Н. – М., 2003.
5. Язык СМИ как объект междисциплинарного исследования. Часть 2 / Отв. ред. Володина М.Н. – М., 2004.
6. .Паршин П.Б. Исследовательские практики, предмет и методы политической лингвистики // Проблемы прикладной лингвистики – 2001 / Отв. ред. Новиков А.И. – М., 2001.
7. Зевахина Т.С. От параметров слова к параметрам текста. – М., 2002.
8. Городецкий Б.Ю. Актуальные проблемы прикладной лингвистики // Новое в зарубежной лингвистике. Вып. XII. М., 1983.