Е.В. ПОЛУХИНА, Д.В. ПРОСЯНЮК

МЕТОДЫ АНАЛИЗА ТЕКСТА В СМЕШАННОМ[1] ДИЗАЙНЕ ИССЛЕДОВАНИЯ[2]

Смешанный дизайн как современная исследовательская практика

Стратегия «смешивания» методов (mixed methods research) [Tashakkori, Teddlie, 1998; Creswell, 2003; Morgan, 2013] получила широкое распространение в западной практике [Полухина, Савинская, 2014]. В самом общем виде – это «способ проведения исследования, в котором сочетаются элементы качественных и количественных исследовательских подходов (оптики, процедур сбора и анализа данных, методов и пр.) для всестороннего и глубокого анализа и решения широкого круга задач» [Burke, Onwuegbuzie, 2004, p. 15]. Современные отечественные исследователи осознают необходимость данного подхода и отмечают, что «…поворот социологии к текстуальности и визуальности способствует увеличению множественности и разнородности данных о социальном мире, что, соответственно, должно вести к многовидовому подходу (*микс-методу*) (курсив авт.) при изучении социальных феноменов, что видится как интеграция или одновременное использование качественной и количественной стратегии при изучении социальных объектов» [Семенова, 2014, с. 8].

Основные характеристики традиционных количественных исследований – дедуктивный подход, подтверждение теорий / гипотез, стандартизированный, формализованный сбор данных и статистический анализ. Качественные исследования же опираются на индуктивную логику, непрерывный поиск, разведку, выдвижение гипотез / теорий, неформализованность, неразрывность сбора и анализа данных. Исследователь выступает в качестве основного инструмента, где его субъективность выступает основой познавательных возможностей.

Принципиальным моментом «смешанного» дизайна исследования является комбинация сильных сторон каждого из методов. Таким образом, результаты исследования будут превосходить данные, полученные с помощью одного из методов.

Гносеологической платформой стратегии смешивания методов является *классический прагматизм*, представленный, в первую очередь, идеями Ч. Пирса, У. Джеймса и Дж. Дьюи. Прагматизм, как версия позитивизма, возник в Америке в конце XIX в. Прагматизм позволяет понять, каким образом может быть достигнута наибольшая эффективность, полезность от объединения различных способов действия. Приверженцы школы прагматизма полагают, что научные методы должны применяться для поиска ответа, «наилучшего» из всех возможных. Сторонники школы стремятся рассматривать различные точки зрения, перспективы и взгляды, используя оптику качественных и количественных исследований. Прагматический подход предполагает взаимный интерес качест-

[1] В российской практике сочетание качественного и количественного подходов называют «комплексными» исследованиями. Мы хотим внести корректировку в понятийное обозначение этого направления исследований, предложив термин, согласующийся с международной практикой. Таким образом, слово «комплексный», скорее, обозначает сложность, многокомпонентность составных частей исследования, а не идею сочетания методов. Понятие «смешивание» представляется адекватным с точки зрения деятельности, которая происходит в исследовательском процессе. Термин «смешивание» подходит, во-первых, когда необходимо формально соединить методы в одном исследовании, оставив сущностную уникальность качественного или количественного характера исследования, следуя прагматическому подходу; во-вторых, «перемешать до однородности», интегрировать, создать полуформализованный, полустандартизованный метод, находящийся на середине методического континуума (по степени формализации / стандартизации).

[2] Статья написана в рамках научного проекта «Смешанные стратегии исследования: потребности, дизайн и процедуры воплощения» (№ 15-05-0012), выполненного при поддержке Программы «Научный фонд НИУ ВШЭ» в 2015 г.

венного и количественного подходов. Его не интересуют природа изучаемой реальности и вопросы «истины». Прагматизм подчеркивает приоритетность действия как основы знания, где акцент делается, скорее, на *полезности* получаемого знания. С прагматической точки зрения исследование является одной из форм действия по достижению цели, в основе которой лежит решение исследовательских задач [Morgan, 2013, p. 42–43].

Методы анализа текста: Формализованный vs неформализованный подход

Контент-анализ является главным методом текстового анализа. Он выступает одним из наиболее интересных методов в социальных науках. Контент-анализ – это метод как сбора данных, так и анализа содержания текста, который может быть реализован в рамках качественного («неформализованного») и количественного («формализованного») подходов. Слово «контент» («содержание» – в переводе с англ.) относится к широкому кругу данных – словам, рисункам, символам, которые могут быть объектом коммуникации. Слово «текст» предполагает определенные рамки и означает нечто написанное, видимое или произнесенное, которое выступает как пространство коммуникации и репрезентации. Это пространство может включать в себя книги, газетные или журнальные статьи, объявления, выступления, официальные документы, кино- и видеозаписи, песни, фотографии, этикетки или произведения искусства [Ньюман, 1998].

Контент-анализ используется как исследовательский инструмент более 100 лет. Сфера его применения междисциплинарна и включает такие науки, как история, журналистика, политические науки, психология и т.д. Так, на первом заседании Германского социологического общества в 1910 г. Макс Вебер предлагал использовать его для анализа газетных текстов [Krippendorf, 1980].

Спектр современных методов формализованного анализа текстов значительно расширился и помимо контент-анализа включает кластерный анализ, тематическое моделирование, анализ тональности и пр. В нашей работе формализованный подход представлен методом кластерного анализа, базирующегося на

представлении текста в виде «мешка слов» (bag of words). Основные допущения данного подхода: а) порядок следования слов / документов в корпусе не имеет значения в тексте; б) текст рассматривается как неупорядоченная совокупность слов (вектор, состоящий из частот слов), где каждое слово имеет равный «вес»; в) слова, встречающиеся часто, исключаются из анализа; г) разные формы слов приравниваются к одному значению.

К формализованным подходам также относится анализ тональности, который нацелен на выявление эмоциональной «окраски» текста. Этот подход разрабатывается в работах Б. Лью, Б. Панга, Л. Ли и С. Вайтинатан и др. [Bing, 2012]. Преимущество формализованного подхода к анализу текста состоит в возможности обработки больших массивов. Не менее важной чертой подхода является «объективность» кодирования – нивелирование субъективности исследователя на этапе обработки данных. Основное ограничение – технические возможности компьютеров, а также учет «прямого» значения слов, безразличие к жанрам, скрытым смыслам, коннотациям. Один из недостатков формализованного подхода – определение темы как совокупности слов в тексте, однако их семантика не эксплицирована. Этот недостаток призван компенсировать альтернативный метод – неформализованный анализ, где процедуру кодирования совершает исследователь «вручную», самостоятельно («эвристическое» кодирование).

Неформализованный анализ рассматривает текст как совокупность смыслов. Текст трактуется как авторское описание, реализуемое с помощью целенаправленного конструирования значений. Исследователя интересуют выявление и толкование смыслов, явно и неявно транслируемых автором, а также интерпретация, реконструкция позиций и типов аргументации, способ преподнесения и видение автором социальной реальности. Этот анализ вытекает из теории аргументации [Attride-Stirling, 2005], он основан на индуктивном подходе, имеет описательный характер и решает поисковые аналитические задачи [Guest, MacQueen, Namey, 2012]. Неформализованный анализ требует активного участия со стороны исследователя. Этот подход выходит за рамки подсчета слов / фраз и сосредоточивается на выявлении и описании значений. В результате такого анализа происходит определение семантической структуры текста. Исследователем разрабатываются коды – «маркеры»

тем, используемые в дальнейшем анализе. Существуют две точки зрения на сущность этого анализа и его соотношение с другими подходами к анализу текста. Ряд исследователей [Braun, Clarke, 2006] полагают, что *неформализованный* (или «тематический», «качественный» контент-анализ) *анализ является интегральным методом*: включает процедуры, заимствованные у таких методов, как обоснованная теория, дискурс-анализ и др. Метод перенимает преимущества других в рамках теоретического и методологического арсенала и адаптирован к прикладным исследованиям. С другой стороны, некоторые авторы отмечают [Boyatzis, 1998], что *данный вид анализа* не является самодостаточным методом, а, скорее, *представлен алгоритмом – кодированием*, который используется также и другими методами.

Образ современной России в текстах газеты «Нью-Йорк таймс»: Процедура работы с текстами

Цель проведенного эмпирического исследования – выявление и описание образа Российской Федерации в американском издании – «Нью-Йорк таймс». Политические и социальные события последних десятилетий, динамика взаимоотношений двух «империй» свидетельствуют об актуальности вопросов, связанных с изучением образа РФ в американских средствах массовой информации.

Эмпирическую базу исследования составляет корпус статей «Нью-Йорк таймс» о России за период августа 2011 г. – июля 2012 г. В это время уровень информационного внимания к событиям в России был достаточно высок, так как проходили думские и президентские выборы, а также был назначен новый состав кабинета министров. Пики публикационной активности приходятся на периоды думских и президентских выборов декабря 2011 г. – марта 2012 г. В исследовании были использованы различные методы анализа текста в зависимости от степени формализации – формализованные (кластерный анализ и сентимент-анализ (метод определения тональности текста) и неформализованный (тематический) анализ.

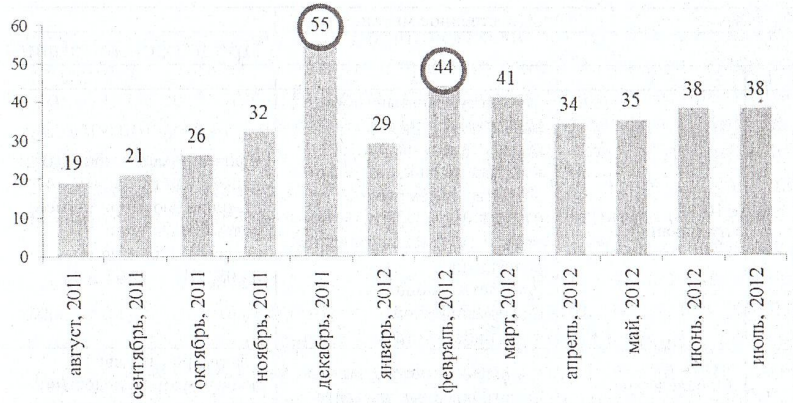

Рис.
Динамика количества публикаций о России в издании «Нью-Йорк таймс», август 2011 г. – июль 2012 г.

Было сформулировано рабочее определение образа России (интегральный конструкт, состоящий из совокупности характеристик России, транслируемых СМИ). Каждый элемент образа соотносится с темой, в контексте которой описывается Россия. Для индуктивного выделения и описания интегрального образа России было необходимо: а) определение тематической структуры и профиля изучаемых текстов; б) последующий анализ каждой из тем, сфокусированный на характеристиках России [Просянюк, 2014]. Алгоритм проведения исследования включает шесть последовательных этапов (см. табл. 1).

Таблица 1

Алгоритм проведения исследования

№	Наименование этапа	Описание этапа	Результат реализации этапа
1	2	3	4
1	Определение источника (источников)	Источником был выбран «Нью-Йорк таймс» – качественное издание, наиболее предпочитаемое элитой и имеющее значительное влияние на	Выбран источник материалов исследования – «Нью-Йорк таймс»

		общественное мнение	

Продолжение таблицы

1	2	3	4
2	Отбор публикаций	Многоступенчатый отбор. На первом этапе отбора была осуществлена сплошная выборка публикаций «Нью-Йорк таймс» за период 01.08.20011–31.07.2012. На втором этапе были отобраны релевантные (удовлетворяющие информационной потребности) статьи	Сформирована выборка исследования. Размер – 411 статей «Нью-Йорк таймс», релевантных теме исследования. Период – 01.08.2011 – 31.07.2012
3	Определение единиц анализа	Разделение текста статьи на два структурных элемента – заголовок и текст	Сформированы две подвыборки исследования – заголовки статей и тексты статей
4	Определение тональности заголовков: автоматическое и эвристическое	Проведение автоматической и эвристической оценки тональности заголовков статей	Выявлена негативная тональность заголовков. Единственной темой, освещаемой исключительно положительно, является российская культура
5	Классификация текстов статей: автоматическая и эвристическая	Проведение автоматического и эвристического тематического анализа текстов статей («bottom-upapproach»). Автоматический тематический анализ проводился двумя методами: кластерный анализ и тематическое моделирование	Выявлено четыре кластера с помощью кластерного анализа и 11 кластеров эвристическим кодированием
6	Выявление и описание элементов образа России на основе проведенной классификации	Индуктивное выделение и описание элементов образа России на основе выделенных эвристическим кодированием тем	Выявлен и описан интегральный образ России, состоящий из следующих элементов: внутренняя политика России, выборы, протесты; внешняя политика России; международные экономические отношения России; культура России; прочие темы (менее 1% корпуса)

Этап 1. Определение источника / источников текстовых данных. Эмпирическим источником был выбран «Нью-Йорк

таймс» – издание, наиболее предпочитаемое элитой и имеющее значительное влияние на общественное мнение. Газета уделяет особое внимание иностранным новостям. Издание «Нью-Йорк таймс» публикует как «жесткие» (критические / политические), так и «мягкие» (нейтральные) новости. «Нью-Йорк таймс» традиционно относится к «качественной» прессе как ориентированной на аудиторию с высоким экономическим и образовательным статусом.

Этап 2. Отбор публикаций. При работе с большими корпусами текстов начальный этап – отбор источников – исключительно важен. Результат, полученный на данном этапе, предопределяет дальнейший ход исследования и качество данных. Основные способы отбора текстов соотносятся с типами выборок и достаточно разработаны в контент-анализе [Krippendorff, 1980; Krippendorff, 2004].

В ходе исследования был использован многоступенчатый отбор, включающий процедуры сплошной выборки[1] (отобрана 2041 статья) и последующего экспертного отбора релевантных статей[2] (411 статей). Фактически оказалось, что 80% исходной выборки (формированных с помощью указания ключевых слов) является «шумом» для настоящего исследования. Полученный результат свидетельствует о необходимости верификации корпуса текстов исследования, отобранных с помощью применения формализованных методов. Соотношение нерелевантных и релевантных статей 100 к 20% (5/1) соответственно свидетельствует, что интеграция

[1] Публикации «Нью-Йорк таймс» за период 01.08.20011 – 31.07.2012 по всем разделам («Политика», «Общество», «Культура» и пр.). Поиск проводился по ключевым словам «Russia», «Russian Federation». Поиск осуществлялся с помощью информационной базы данных LexisNexis.

[2] При первом ознакомлении с материалом стало очевидно, что далеко не все статьи отвечают задачам исследования. Так, статья может содержать ключевое слово «Russia» единожды – в расшифровке аббревиатуры BRIC. Данная статья формально является релевантной, но не отвечает информационной потребности исследователя, так как не содержит в себе эксплицированных в тексте индикаторов предмета исследования – мнений, оценок России и пр. С другой стороны, статья может быть посвящена нарушению прав человека в российском обществе, т.е. очевидным образом соответствовать информационной потребности исследователя, но также содержать ключевое слово «Russia» в единственной фразе. Такие статьи будем называть «пертинентными» (релевантыми, т.е. соответствующими информационной потребности).

формализованных и неформализованных процедур отбора является приоритетной.

Этап 3. Определение единицы анализа. Определение единиц анализа при работе с большими массивами текстов является неоднозначным вопросом. С одной стороны, в сложившейся практике принято *отождествлять текстовый массив и единицу анализа*. Процедура рассмотрения текста как единицы анализа отвечает прагматическим соображениям – при работе с корпусами размером в несколько сотен / тысяч документов разделение текстов на единицы анализа трудозатрано. С другой стороны, лингвистические дисциплины рассматривают газетные *статьи как особый вид текста*. Следуя за Т. Ван Дейком, особого внимания заслуживает *анализ как отдельных структурных элементов, так и макроструктурных характеристик* (тематической структуры текста). У каждого структурного элемента статьи свои социально-коммуникативная функция, строение, лексико-семантические особенности. Ввиду специфики каждого элемента / сегмента текста статьи целесообразно, на наш взгляд, рассматривать их как отдельные единицы анализа. Применительно к нашему исследованию отдельные элементы формирования образа России содержательно должны рассматриваться в комплексе.

В журналистике подробно изучена структурная организация текстов разных жанров. М.Н. Ким [Ким, 2001] отмечает, что в аналитической статье главная роль отводится: а) выдвижению основного тезиса для доказательства; б) построению системы аргументации, раскрывающей суть выдвинутого тезиса; в) выводам из системы доказательства [ibid]. Данные структурные элементы создают трехчленную структуру статьи: начало (основной тезис) – основная часть (аргументы) – заключение (вывод). Как правило, основная цель аналитической статьи заключается в демонстрации различных точек зрения, мнений, оценок. Представляется возможным использовать укрупненные единицы, а именно *две основные части статьи*, имеющие принципиально различные цели и структуру: заголовок (для привлечения внимания, описания основной идеи статьи) и «тело» статьи (включая заключение).

Этап 4. Определение тональности заголовков: автоматическое и неавтоматическое («ручное»). СМИ используют как средства рационального (аргументированного), так и эмоционального

воздействия. Было решено определять тональность заголовков (уровень предложений) по нескольким причинам. Во-первых, заголовок представляет главную тему статьи. Как правило, тема в заголовке традиционно одна. Последнее дает основание полагать, что возможно адекватно определить именно тональность заголовка. Во-вторых, текст статьи, как правило, содержит спектр тем и аргументов, следовательно, однозначно определить его тональность затруднительно. Напомним, что метод оценки тональности исходит из допущения, что каждый документ выражает мнение об одном объекте. Таким образом, анализ не применим к документам, которые оценивают или сравнивают несколько объектов. В-третьих, специфика новостного жанра допускает, что получатель сообщения в ряде случаев ограничивается прочтением заголовка из всего текста новости.

Тональность текста может быть определена двумя альтернативными методами: автоматическим (формализованным) либо неавтоматическим (тематическим кодированием). Для формализованного определения тональности использовалось онлайн-программное обеспечение Tweenator[1]. Кодировщик определял тональность по шкале «Негативно» – «Нейтрально» – «Положительно». На вход программному обеспечению были представлены заголовки, которые были определены кодировщиком как «не нейтральные».

Адекватно тональность заголовка была определена в 55% случаев. Типичные ошибки соответствуют ожидаемым проблемным зонам формализованного анализа тональности (нечувствительность к иронии, сарказму), что, в свою очередь, и приводит к смысловому искажению («смыслы плывут»). Подход «обучения с учителем» (обучается машинный классификатор на заранее размеченных текстах, а затем используют полученную модель при анализе новых документов) имеет явные недостатки, которые негативно повлияли на результат. Полученные параметры являются «черным ящиком», не подлежат дальнейшей настройке и корректировке. Отсюда следует,

[1] Данное программное обеспечение разработано для определения тональности на уровне предложений таких элементарных контекстов, как «твиты» (сообщения из социальной сети Twitter) или заголовки новостных статей. Программа основана на методе обучения с учителем (machine learning approach). Обучающей выборкой были данные Stanford Twitter Sentiment Data, которые были собраны в период с 6 апреля по 25 июня 2009 г. Выборка состояла из 1,6 млн твитов и такого же количества твитов, содержащих эмотиконы (способы выражения эмоций).

что результат применения «обученного» алгоритма на новой выборке требует обязательной экспертной верификации («ручной» проверки). Таким образом, применение тематического («ручного» / неформализованного) кодирования значительно повышает качество формализованной оценки тональности текста.

Проведенный анализ тональности заголовков показал, что новостные материалы, посвященные России, носят преимущественно негативную эмоциональную окраску. Анализ тем показывает, что единственной темой, освещаемой положительно, является российская культура (в большинстве случаев – балет).

Этап 5. Классификация текстов статей: автоматическая и эвристическая. Элементы образа России могут быть индуктивно определены на основе анализа тем, представленных в текстах «Нью-Йорк таймс», где упоминалась Россия. Для решения данной задачи был проведен автоматический (формализованный)[1] и тематический (неформализованный) анализ текстов статей [Braun, 2006]. Автоматически в корпусе статей было выделено четыре кластера: Krymsk (19%), Oil (13%), Putin (40%) и Syria (28%). Результаты проведения неформализованного тематического анализа представлены в табл. 2. В данном случае автоматически удалось идентифицировать две самые частотные темы (как количественно, так и содержательно). Вместе с тем по наполнению кластеры, выделенные автоматическим путем, гетерогенны дальнейшему анализу и интерпретации не подлежат.

Таблица 2

Кластеры текстов статей, полученные с помощью неформализованного / тематического анализа

№	Тема	Доля (в %)
1	Внутренняя политика России (выборы, протесты, деятельность правительства и пр.)	51
2	Внешняя политика России	33
3	Культура России	4
4	Экономическая политика России	4
5	Нарушение прав человека в России	2
6	Возможности организации бизнеса с Россией	2
7	Остальные темы (менее 1% каждая)	4

[1] Формализованный тематический анализ проводился методом кластерного анализа, алгоритм двукластерного решения (bisectingk-means), косинусная мера. Использовалось программное обеспечение TLab.

Этап 6. Выявление и описание элементов образа России. Каждому сообщению соответствовала не одна, а несколько тем. Поэтому описание основных тем находится не в четком соответствии с подкорпусом информационных сообщений.

Заключение

Для изучения такого явления, как образ страны, сочетание формализованных и неформализованных подходов к анализу текстов является необходимым и естественным. Технические средства изменяют структуру текста, могут приводить к искажению его первоначального значения. Так, отобранный для анализа текстовый фрагмент, помещенный в программное обеспечение, претерпевает существенные изменения. Автор повествования сменяется автором обработки и анализа. В поле исследователя остаются преимущественно образованные коды, к которым обращены фрагменты текста.

Важность учета контекстуальности текстовых данных делает необходимым «индивидуальную» работу, где программное обеспечение представлено в качестве оптимизирующего механизма. Так, авторами в целях изучения формируемого в СМИ образа страны реализован смешанный дизайн исследования, интегрирующий как формализованные, так и неформализованные методы анализа текста. Таким образом, преимуществом смешанного дизайна исследования является взаимообогащение познавательных возможностей, данных и интерпретаций.

Список литературы

Ньюман Л. Неопросные методы исследования // Социологические исследования. – М., 1998. – № 6. – С. 119–129.

Полухина Е.В., Савинская О.Б. Через исследования к доверительному глобальному сотрудничеству: перспективы качественных исследований в XXI веке // Социологические исследования. – М., 2014. – № 1. – С. 122–125.

Просянюк Д.В. Образ России в призме социально-проектных и информационных технологий // Власть. – М., 2014. – № 1. – С. 50–54.

Семенова В.В. Стратегия комбинации качественного и количественного подходов при изучении поколений // Интер. – 2014. – № 8. – С. 5–15.

Attride-Stirling J. Thematic networks: ananalytic tool for qualitative research // Qualitative research. – 2001. – N 1. – P. 385–405.

Bing L. Sentiment analysis and opinion mining. – L.: Morgan & Claypool publish, 2012. – Mode of access: http://www.dcc.ufrj.br/~valeriab/DTM-SentimentAnalysis AndOpinionMining-BingLiu.pdf (Дата посещения: 12.02.2015.)

Boyatzis R.E. Transforming qualitative information: thematic analysis and code development. – Thousand Oaks, CA: Sage. 1998. – 184 p.

Braun V., Clarke V. Using thematic analysis in psychology //Qualitative research in psychology. – 2006. – Vol. 3, N 2. – P. 77–101. – Mode of access: http://eprints.uwe.ac.uk/11735/2/thematic_analysis_revised (Дата посещения: 12.02.2015.)

Burke J.R., Onwuegbuzie A.J. Mixed methods research: A research paradigm whose time has come // Educational researcher. – 2004. – Vol. 33(7). – P. 14–26.

Creswell J.W. Research design: qualitative, quantitative, and mixed approaches. – Thousand Oaks. CA: Sage, 2003. – 273 p.

Guest G., MacQueen K., Namey E. Applied thematic analysis. – Thousand Oaks, CA: Sage, 2012. – 320 p.

Krippendorf K. Content analysis: An introduction to its methodology. – Beverly Hills. Thousand Oaks, CA: Sage, 1980. – 188 p.

Krippendorff K. Content analysis: An introduction to its methodology. – Thousand Oaks, CA: Sage, 2004. – 413 p.

Morgan D. Integrating qualitative and quantitative methods: A pragmatic approach. Sage publications, 2013. – 288 p.

Tashakkori A., Teddlie C. Mixed methodology: combining qualitative and quantitative approaches. – Thousand Oaks, CA: Sage. 1998. – Vol. 46. – 200 p. [Applied social research methods series.]