# КОНЦЕПТУАЛЬНОЕ ОБОСНОВАНИЕ КОМПЬЮТЕРНОГО АНАЛИЗА МАССИВОВ С ТЕКСТАМИ[1]

Е.А. Каневский, Г.И. Саганенко

*(Санкт-Петербург)*

Опыт социологов в применении текстов как баз эмпирической информации в целом достаточно ограниченный. Это использование простейших открытых вопросов, формализованный анализ текстов СМИ, качественный и в общем всегда трудоемкий метод получения и анализа текстов интервью, биографий. Как правило, все варианты обходятся без серьезного компьютерного текст-анализа по существу.

Рассматриваемая нами задача состояла в том, чтобы предложить соответствующую, гибкую и разнообразную, компьютерную систему поддержки исследований с текстами, с тем, чтобы социологи чаще и с большей легкостью выходили на исследования с массивами текстов, чтобы при работе со сложными текстами могли значительную часть своих аналитических проблем решить с помощью возможностей такой системы.

*Ключевые слова:* базовые фразы, вторичный признак, дерево имен, дерево частот, группа фраз, идентификация, итеративная классификация, класс фраз, классификатор, классификация, ключевое слово, контент-анализ, метод открытых вопросов, нормативные словари, открытый вопрос, сравнение текстовых подмассивов, текст-анализ, фраза.

## Методологическая ситуация в данной области

1) В настоящее время в самых разных областях гуманитарных и социальных исследований растет интерес к "мягким" исследовательским концепциям, "мягким" технологиям и измерениям [1, 2, 3, 4]. Новые концептуальные подходы, базы эмпирических данных с принципиально новыми параметрами, новые компьютерные технологии встраиваются друг в друга, обеспечивая кумулятивный эффект.

Однако две эффективные системы получения эмпирического знания в социологии: методология массовых обследований по жестким стандартизированным методикам и "мягкие" технологии - в целом функционируют на непересекающихся исследовательских полях, методологическая парадигма социологического знания не расширяется за счет их дополнительного освоения. Что касается отечественной социологии, то в методологических работах (которых, кстати, очень мало) имеется явный крен на осмысление лишь формализованной ситуации массовых опросов [5, 6, 7, 8].

2) Специфические познавательные возможности содержатся в текстовых массивах, циркулирующих в огромном количестве в обществе. Разные области науки обращаются к поиску и анализу текстовых материалов. В частности, социология также обращается к анализу разного рода текстов. Она анализирует наличные тексты, в том числе материалы средств массовой информации, политические материалы в виде программ партий и кандидатов в электоральных кампаниях, уставы партий и движений, биографии и дневники, научные публикации и др. Социология сама стимулирует появление в обществе специальных текстов, проводя конкурсы сочинений и автобиографий, организуя интервью, используя "открытые вопросы".

Проблематика текст-анализа представляется междисциплинарной областью, имеющей множество приложений. Технологии анализа текстов отличаются большим разно-

[1] Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований. Грант №96-0680216.

образием: это сплошной и частичный анализ текстов, качественный и количественный анализ.

Систематический анализ текстового содержания часто называют контент-анализом и понимают его по-разному: от формального стандартизированного метода из области "жесткой" количественной технологии до пластического метода качественного подхода и даже универсальной концепции исследовательского поиска.

3) Что касается контент-анализа как актуального метода в отечественной социологии, - то он за многие годы имел лишь две базированные "репрезентации" - две конференции и соответствующие им сборники тезисов [9, 10]. Однако ни одной монографической работы контент-анализу не было посвящено. В отдельных публикациях принципиального развития метод не получает, и в статьях методологического характера он рассматривается, в основном, как стандартизированная техника для анализа содержания газет [11]. За два десятка лет появилось не более полудюжины статей, заслуживающих определенного внимания.

С учебной литературой ситуация еще менее утешительная: там либо вообще нет упоминаний о методе, либо есть отдельные абзацы с нечеткой презентацией стандартизированного метода.

В практике эмпирических исследований, как правило, контент-анализ видится сугубо формализованным методом, типичным примером является методика Б.Грушина и Л.Федотовой, примененная при контент-анализе публикаций газет по экологическим проблемам [6].

4) Прямое отношение к нашей теме имеет проблематика использования открытых вопросов. Методологическое ее рассмотрение хоть и заслужило большего внимания со стороны зарубежных ученых, однако дается очень узко и касается двух-трех аспектов проблемы, где в первую очередь обсуждается "активность" респондента при ответах на закрытые и открытые вопросы. Проблемы анализа информации видятся примерно такими же, как и в отечественном контент-анализе [7, 12]. Методологическое осмысление метода открытых вопросов, судя по российским и частично зарубежным учебникам, проводится очень схематично и весьма узко, с аргументацией, почерпнутой из примитивных примеров, и даже содержит принципиальные ошибки. Практическое же использование открытых вопросов и связанной с ним методологии и процедур контент-анализа проводится в ограниченном количестве и в целом не удовлетворительно.

На самом деле имеется существенная специфика тех исследовательских ситуаций, где метод открытых вопросов наиболее релевантен и имеет высокие разрешающие возможности, дает более адекватные и надежные результаты. Нам представляется очень важным проанализировать такие ситуации и предложить их перечень, дать их обоснование.

В частности, познавательные возможности метода открытых вопросов, с нашей точки зрения, успешно реализуются в следующих ситуациях:

- исследование направлено на весьма широкий объект;
- исследование осуществляется в период социальных трансформаций;
- исследование панельное, тем более если оно осуществляется в период социальных трансформаций;
- сравниваются разные социально-культурные ситуации и, в частности, разные социальные группы, структуры их ценностей;
- "респондент" в некоторой сфере "квалифицированней" исследователя;
- исследование готовится по "жесткой" анкете (тем более исследование массовое и ответственное) и, соответственно, проводится отладка инструментария для грамотного отбора и формулирования вопросов и формализованных ответов.

5) Что касается ситуации с использованием компьютеров для анализа текстов, то здесь обнаруживается несколько принципиально разных подходов. В частности, относительно часто за рамками компьютерных технологий сначала осуществляется формализованное описание документов: разрабатывается специальная априорная жесткая методика формализованного кодирования текстового документа, "вручную" осуществляется

описание каждого документа (получается своего рода "библиографическое описание" документа) и тем самым создается массив формализованных цифровых описаний. А затем уже привлекаются компьютеры для анализа такого массива как совокупности простых стандартных анкет для получения частотных распределений, долей, средних и пр.

Компьютеры используют для промежуточных вспомогательных операций, для реализации заранее расписанного стандартного описания теста.

И, наконец, кодирование, классификация, индексирование полных текстов или их фрагментов, манипуляция с текстами, их анализ, полностью осуществляются в рамках компьютерной технологии.

Что касается отечественной ситуации, то самостоятельного интереса к компьютерным технологиям исследователи не проявляют. Они либо пользуются теми процедурами работы с текстами, которые поддерживаются любой современной базой данных (например, это получение алфавитных словарей с указанием частот), либо осуществляют уже упомянутую статистическую обработку текстовых документов, обработанных предварительно "вручную".

За рубежом ситуация в целом другая и, в частности, проблематика анализа текстов рассматривается как междисциплинарная область, имеющая разнообразные приложения от археологии до дискурсного анализа.

Например, с 1972 г. существует общество **SCCAC** (Society for Conceptual and Content Analysis by Computer), объединяющее людей, имеющих интересы в области анализа текстов. Его основными задачами являются проблемы систематической организации и выявления знания, связанного с большими массивами текстов и базами данных. В проблематику включают все то, что относится к анализу естественного языка, концептуальной классификации, методам лексикографии, индексирования, текст-анализа и др. Общество выпускает информационный бюллетень, участвует в подготовке сборников и монографий.

Проблематика текстового анализа и ее компьютерные решения обсуждаются на представительных конференциях, в частности, на конференциях 1995 г.: "Визуализация категориальных данных", Кельн и "Текст-анализ и компьютеры", Маннгейм.

Известен ряд компьютерных программ по анализу текстов: **PLCA, KWALITAN, CETA, AQUAD, WINMAX, TEXTPACK PC, TACT** [4, 13]. Их рассмотрение дает возможность понять современные проблемы и пути поиска решений в данной сфере.

В частности, **KWALITAN** (Нидерланды) позволяет создавать архивы документов, осуществлять поиск и выборку полных документов или их частей, разделять тексты на любое количество сегментов и содержание каждого сегмента описывать совокупностью кодов, просматривать коды в алфавитном порядке вместе с их частотами или упорядочивать коды в соответствии с их частотами. Программа позволяет выделять конкретные слова, включать их в отдельные списки и получать для них частотные распределения. Для выделенных слов могут быть выведены все контексты, в которых имеется данное слово. Сегменты выделяются путем кодов и их комбинаций с помощью логических операторов.

Широко применяемая ныне программа **TEXTPACK PC** (разрабатывается в центре ZUMA, Германия, с 1972 г., в настоящее время имеется уже 5-я версия) обладает компьютерными возможностями в таких областях как контент-анализ, литературный и лингвистический компьютерный анализ [14]. Как основные достоинства программы разработчики отмечают широкие возможности для вычисления частот выделенных слов и извлечения ключевых слов с их контекстом. Другой важной характеристикой программы считается наличие специальной целевой процедуры, которая позволяет категоризировать, классифицировать, вводить переменные для любого рода текстов согласно так называемым "контент-аналитическим словарям". Существует множество сервисных процедур, которые помогают просматривать, распечатывать текстовые файлы, разрабатывать и обосновывать словари, и, что наиболее важно, связывать итоговые цифровые выводы, то есть частоты

категорий, со статистическими пакетами, такими как **SPSS** или **SAS**, для дальнейшего количественного и логического анализа.

Существует ряд программ, позволяющих изучать логические и семантические структуры текстов, сводя их к матрицам или сеткам через введение кодов для объектов и обозначение связей между ними.

Как видно, большинство систем анализа содержания текстов в качестве элемента содержания (единицы анализа) использует слово. Результаты такого анализа оказываются формальными, не проясняющими содержание по существу, поскольку отдельное слово, как правило, не отражает мысль автора, одного слова недостаточно для понимания смысла высказывания или фрагмента текста.

При всем том что ведется достаточно интенсивная работа по разработке компьютерной поддержки этапов получения первичной текстовой информации и облегчения последующей работы с текстовыми массивами, нельзя пока сказать, что исследования с текстами широко используются в социологии. Они все еще весьма трудоемки, занимают значительное время, требуют от исследователя серьезного опыта. Еще более скромная ситуация наблюдается в отечественной социологии.

Разработанная в Санкт-Петербурге система **ДИСКАНТ** (Диалоговая Итерационная Система Классификации и Анализа Текстов)[1] показывает, что можно в значительной степени снять трудоемкость и повысить оперативность исследований с текстовыми массивами, более того, можно добиться доказательности получаемых в социологии результатов и выводов.

Сравнение характеристик обсуждаемой системы с аналогичными разработками, представленными на рынке компьютерных технологий для социальных исследований, показывает, что в системе **ДИСКАНТ** имеется ряд весьма серьезных достоинств, интересных и эффективных решений. В частности, это показала дискуссия при демонстрации системы **ДИСКАНТ** на конференции "Текст-анализ и компьютеры" [4], а также в рамках доклада [15] Маннгейм, Германия, 1995.

## Основные особенности системы ДИСКАНТ

*2.1. Классификация текстов в итеративном режиме*

Система ДИСКАНТ предназначена для обработки анкетной информации "мягкого" и "жесткого" типа, а именно, произвольных текстовых ответов на систему открытых и полузакрытых вопросов анкеты и "жестких" ответов на формализованные закрытые вопросы (номинальные, квантованные, количественные, многоальтернативные), а также для совместной обработки информации обоих типов.

Основными и наиболее эффективными возможностями системы, на наш взгляд, являются итеративная классификация текстовых единиц (фраз), анализ результатов классификации, сравнение классификации для разных подмассивов и даже разных исследований. Рассмотрим эти возможности подробнее.

1) Решение задач классификации и сопутствующих ей процедур осуществляется в итеративном режиме, порциями, что должно позволить получать чистые и весьма удовлетворительные результаты содержательной классификации высказываний.

[1]Методология классификации текстов в итеративном режиме разработана ведущим научным сотрудником Института социологии РАН, д.социол.наук, проф. Г.И.Саганенко (Санкт-Петербург, 7-я Красноармейская 14/25, ИСАН); принципы идентификации фраз и построения базы данных для социологических исследований разработаны старшим научным сотрудником Санкт-Петербургского экономико-математического института РАН к.т.н., ст.н.с. Е.А.Каневским (Санкт-Петербург, Чайковского 1, СПбЭМИ РАН), под руководством которого выполнено программирование системы ДИСКАНТ.

Итеративные возможности классификационной процедуры состоят в том, что два основных результата (разработка классификатора и использование его при анализе высказываний) являются в целом итогом полного освоения текста исследователем. По мере осознания им специфики материала, вплоть до самого заключительного момента исследования, открываются новые детали содержательных и формальных характеристик материала. Исследователь имеет возможность учитывать особенности, корректировать результаты на любой стадии.

При этом допустимы разные варианты итераций (варианты опробованы и в принципе позволяют рассчитывать показатели "скорости сходимости" процедур). Например, для первой итерации классификации можно выбрать наиболее распространенные или более определенные и понятные темы, затем обработать следующую порцию тем и фраз и т.д. Или сначала выбрать все суждения из первых (например, 50) анкет, для них решать проблему поиска классов и классификацию суждений, затем добавлять новую порцию и т.д.

2) В рамках общего массива информации система позволяет вести обработку целого ряда различных текстовых подмассивов, требующих автономных процедур классификации и анализа.

В рамках каждого подмассива разрабатывается классификатор и в соответствии с ним проводится классификация высказываний. Собственно классификатор отражает структуру, заложенную в содержании текстового подмассива, и, как правило, реализуется в виде "дерева" (как бы "дерева имен"), каждая полная ветвь которого может иметь 1-2 звена-имени. Соответственно структура всех текстовых подмассивов описывается в виде нескольких разных "деревьев".

3) Обработка отдельного текстового массива предполагает идентификацию всех составляющих его единичных фраз в конечном итоге в соответствии с разработанным классификатором, при этом итоговый результат классификации представляет "дерево с весами" - с указанием для каждого "имени" классификатора количества отнесенных к нему суждений (как бы "дерево имен и частот"). Результаты классификации всех текстовых подмассивов представляют соответственно несколько "деревьев с весами".

4) Разработка классификатора и сама классификация осуществляется внутри системы, не требуя априорных схем, а только в соответствии с содержательным наполнением текстового массива, поиска надлежащего структурирования материала и привлечения доказательств в самом материале. При этом для поиска более оптимальных режимов реализации этих двух задач имеется несколько приемов поддержки (в частности, словари и частотные словари фраз, слов и ключевых слов, разные режимы поиска).

5) Система является многоклассификационной, она позволяет независимо друг от друга анализировать множество отдельных текстовых подмассивов с созданием для них собственных классификаторов и классификаций.

Классификационных моделей может быть реализовано столько, сколько единичных текстовых полей (открытых вопросов). Но, как правило, количество моделей меньше числа таких полей, поскольку обычно существуют серии вопросов, направленных на сходную социальную тематику "описываемых" одним классификатором.

6) Система "параллельна" в том смысле, что она позволяет переходить от одного подмассива (в любой момент реализации его классификационной модели) к другому и выполнять требуемые для него операции.

7) Система "диалоговая" в том смысле, что решения о вводе классов в классификаторе, о непосредственной классификации фраз в соответствии с позициями классификатора, об идентификации фраз с оптимальными классифицированными суждениями принимает исследователь, система же только предлагает средства оптимального выбора из множества близких решений.

8) Система проверяема относительно всех ее узловых решений, а при наличии нерелевантности позволяет вносить исправления.

9) Результаты классификации можно сопоставлять для разных текстовых подмассивов и подмассивов респондентов (с предварительным формированием базы "вторичных признаков"). В принципе можно сопоставлять результаты совокупности панельных исследований. В целом это одна из наиболее эффективных и оригинальных процедур системы.

*2.2. Идентификация фраз*

Предлагаемый нами метод содержательного анализа текстов основан на использовании фразы в качестве единицы анализа. "Фраза" - это такой фрагмент полного ответа, который содержит только одно суждение или одну тему, и, как правило, понятен в отрыве от контекста. Фраза может состоять из одного или нескольких слов, а может включать в себя несколько предложений.

1) Разработанный нами процесс контент-аналитического сравнения фраз заключается в следующем. Заранее отбираются и классифицируются некоторые фразы, в дальнейшем называемые базовыми. При классификации каждая такая фраза помещается в соответствующий класс и группу. В каждой базовой фразе желательно выделить ключевые слова (ключи), которые наиболее полно выражают основную мысль данной фразы. На основе базовых фраз создаются три так называемых нормативных словаря: слов, ключей и фраз. Процедура непосредственного сравнения начинается с того, что в нормативном словаре фраз производится поиск исследуемой фразы. Если таковая имеется, то ей присваиваются номер класса и номер группы базовой фразы и на этом процесс сравнения заканчивается.

Если же таковой не оказалось, то есть ни одна базовая фраза полностью не совпала с исследуемой, то производится разложение этой фразы на отдельные слова и анализ каждого слова в отдельности. В начале из нормативного словаря слов производится выборка всех близких к нему слов и сравнение с ним. В случае достаточного совпадения фиксируется номер каждой базовой фразы, к которой относится найденное в нормативном словаре слово. Такая же процедура повторяется со словарем ключей. В процессе анализа подсчитывается количество слов и количество ключей, встретившихся в каждой из зафиксированных базовых фраз. Та из базовых фраз, которая получит максимальный "вес", считается аналогом исследуемой.

2) С целью ускорения процессов идентификации фраз и повышения точности их контент-аналитического сравнения применяются словари "исключаемых" слов и "нестандартных" слов [16]. Первый словарь используется при составлении нормативных словарей слов и ключей и позволяет сократить их размер за счет отбрасывания таких служебных и вспомогательных слов, которые не имеют существенного значения для смысла отдельной фразы. Второй словарь используется непосредственно при сравнении слов и служит для указания размера сравниваемой части слова.

3) Следующим важным моментом при контент-аналитическом сравнении фраз является учет их модальности. Очевидно, что в простейшем случае следует различать два типа фраз: утвердительные и отрицательные. Две фразы, почти полностью совпадающие по своему словарному составу, могут иметь совершенно противоположный смысл, который связан с наличием в одной из них отрицательной частицы или слова (например, "не", "нет", "никак"). При разложении исследуемой фразы на слова и подборе наиболее близкой к ней базовой фразы обязательно сравниваются их модальности, если они не совпадают, то и фразы считаются несовпадающими.

*2.3. О применимости системы **ДИСКАНТ***

1) Система эффективна в применении к анализу текстов, которые изначально имеют определенную структуру, а именно, представляют собой множество суждений, формулировок по актуализации или поименованию некоторых идей, состояний, объектов и т.п. В частности, в социологии такой материал дают суждения экспертов, высказывания по темам

интервью, открытые вопросы (применительно к которым и описана система) и др. В психологии - это текстовые дополнения в методе неоконченных предложений и др. В средствах массовой коммуникации - это названия публикаций и передач, их тематическая направленность и др. В любой науке это архивные описи, библиографические карточки и пр.

2) Кроме того, система в принципе позволяет проводить анализ текстовой неструктурированной информации такого рода как, например, биографии, интервью, предвыборные программы кандидатов или партий, публикации в газетах и журналах и др. В частности, здесь можно первоначально "структурировать" тексты, отбирая подмассивы "фраз" под определенные социальные переменные (dimensions) и далее действовать по описанному алгоритму. Выполнение же для таких текстов частных задач на составление словарей и указателей, подсчет частот, поиск слов и др. не требует никаких серьезных усилий).

3) Система отлаживалась на массивах лонгитюдного исследования, осуществляемого раз в год, начиная с 1989 г. [15, 17, 18]. Основной объем первичной информации получался из ответов на открытые вопросы относительно сложных социальных объектов, таких как "ситуация в обществе", "перспективы общества", "перемены в сфере работы", "значимые характеристики частной жизни". В 1993-1995 г. лонгитюдное исследование получило поддержку фонда Сороса (программа "Research Support Scheme), заключительному отчету по проекту дана высокая оценка экспертным советом Программы RSS.

## ЛИТЕРАТУРА


1. *Ядов В.А.* Социологическое исследование: методология, программа, методы. М.:Наука, 1987.
2. Qualitative Sociology. 1992. Vol. 15. №1.
3. *Strauss A.L.* Qualitative Analysis for Social Scientists. Cambridge Univ. Press, 1991.
4. Text Analysis and Computers//Conference Programme and Abstracts. Mannheim: ZUMA, Center for Survey and Methodology, 1995.
5. *Белановский С.А.* Методика и техника фокусированного интервью. М.:Наука, 1993.
6. *Грушин Б.А., Федотова Л.Н.* Методика анализа содержания прессы//Разработка научных основ формирования экологического сознания населения страны. Ч. 2. М., 1990.
7. *Маслова О.М.* Познавательные возможности метода опроса// Методы сбора информации в социологических исследованиях. Ч. 1. М., 1990.
8. Handbook of Survey Research/Ed. by Peter H. Rossi. Academic Press, INC, 1983.
9. Методологические и методические проблемы контент-анализа. Вып.1-2. М.-Л.:ИСИ АН СССР, 1973.
10. Проблемы контент-анализа в социологии (материалы Сибирского социологического семинара). Новосибирск, 1970.
11. *Коробейников В.С.* Методы формализованного анализа документальных источников//Методы сбора информации в социологических исследованиях. Ч. 2. М., 1990.
12. *Маслова О.М.* Открытые и закрытые вопросы//Методы сбора информации в социологических исследованиях. Ч. 1. М., 1990.
13. *Кошкин О.А.* (О публикации:) *Теш Р.* Качественное исследование: типы анализа и программное обеспечение (Tesch, R. Qualitative Research: Analysis Types & Software Tools. New York: Falmer Press, 1990)//Социологический. журнал, 1994. № 1.
14. TEXTPACK PC. Short Description (by P. Ph. Mohler, C. Zull). Mannheim: ZUMA, 1995.
15. *Saganenko G.* Dynamics of Changes in Russia: Longitudinal from 1989 Investigation with Open Questions and Content-Analyses//Text-Analysis and Computers. Conference Programme and Abstracts. Mannheim: ZUMA, 1995.
16. *Каневский Е.А., Клименко Е.Н., Гайдукова Л.М.* Контент-анализ текстов и проблемы идентификации//Информационные технологии в гуманитарных и общественных науках. СПб: Спб ЭМИ РАН, 1995.
17. *Саганенко Г.И.* Приобретения и потери научной интеллигенции в связи с трансформациями в России//Наука и социальные институты в кризисном обществе. СПб, 1996.

18. *Саганенко Г.И.* Пресса и перестройка глазами научной интеллигенции//Журналист, пресса, аудитория. Вып. 4. Л.: Изд-во ЛГУ, 1991.