УДК 159.98
ББК Ю945

МЕТОДЫ АНАЛИЗА ТЕКСТА: МЕТОДОЛОГИЧЕСКИЕ ОСНОВАНИЯ И ПРОГРАММНАЯ РЕАЛИЗАЦИЯ

О.В. Митина, А.С. Евдокименко

Изложена систематизация представлений о методологических принципах анализа текста и программной реализации уже разработанных методик. Методики анализа текста были систематизированы в 10 групп: интент-анализ, контент-анализ, фоносемантический анализ, дискурс-анализ, нарративный анализ, экспертная оценка текста, графематический анализ, морфологический анализ, синтаксический анализ, семантический анализ. Для каждой из групп приведены примеры программной реализации.

Ключевые слова: текст, методы анализа текста, программы анализа текста.

Расширение контекста применения методов анализа текста, а также поиск зависимостей и связей единиц анализа текста с психическими и социальными процессами создало большое разнообразие подходов и способов исследования, которые на настоящий момент с трудом поддаются систематизации. Развитие компьютерных технологий и увеличение роли информации на настоящий день отвело методам анализа текста особую привилегированную роль. Методы анализа текста применяются при поиске, систематизации, оценке, отборе информации, диагностике, анализе и прогнозировании событий или поведения субъекта, из-за чего широкое применение эти методы получили в системах безопасности. Возможности применения методик анализа теста с интерактивными компьютерными системами и базами данных значительно ускоряют развитие искусственного интеллекта. В связи с этим остановимся на некоторых общих чертах всех методов анализа текста и выделим критерии их систематизации.

Единица анализа текста

Ключевым моментом, который объединяет все методики анализа текста, является то, что в их основе лежат представления о единице анализа. Понятие единиц анализа является крайне важным аспектом, поскольку выступает своего рода аналогом исследуемых (но неконтролируемых) переменных в эксперименте. Под единицами анализа в широком смысле понимаются «такие продукты анализа, которые в отличие от элементов обладают всеми основными свойствами, присущими целому, и которые являются далее не разложимыми живыми частями этого единства...» [5, с. 46–48]. Однако их трактовка в конкретном методе анализа текста может быть принципиально различной. Так, например, под единицей лингвистического анализа текста понимаются инварианты различных лингвистических моделей описания языка (морфема, фонема, предложение, словосочетание, высказывание и др., трактовка которых в различных лингвистических направлениях также различна), соотносящиеся с языком или языковым стандартом. Под единицей психолингвистического анализа понимаются речевые действия и операции, находящиеся друг с другом в иерархических отношениях, которые соотносятся с речевой деятельностью (А.А. Леонтьев), или сегменты сообщения, которые являются функционально оперативными как целые в процессах декодирования и кодирования, поддающиеся уровневому анализу [24, с. 61]. Таким образом, психолингвистическими единицами анализа текста могут выступать [10] фреймы, концепты, когниотипы, эмотемы, пропозиции и др. Единицей же психологического анализа относительно текста выступают компоненты, обеспечивающие речевую деятельность – письмо или произношение, – это обеспечивается тем, что «в высказывании, тексте объективируется вся совокупность психологических условий деятельности и индивидуально-психологические особенности её субъекта» [7, с. 61]. Другими словами, единицами психологического анализа текста могут выступать: представления субъекта о языке, отражение в сознании языковой способности, мотивы, интенции и др. Однако вне зависимости от типа ко всем единицам анализа предъ-

являются два требования, которые обычно трудно совместимы на практике: единицы контент-анализа должны однозначно опознаваться в тексте и одновременно должны быть значимы для последующей (политологической, культурологической, социологической и т. д.) интерпретации.

На интерпретацию конкретных единиц анализа текста и отношения между ними, в свою очередь, влияет модель, основанная как на эмпирических, так и теоретических представлениях о языке и речепорождении. При этом анализ текста может быть представлен как перевод его на метаязык лингвистических, психолингвистических или психологических единиц, или как сопоставление его с теоретической моделью, порождённой этим метаязыком. В связи с этим каждая модель позволяет рассматривать только определенный спектр закономерностей и особенностей текста, допускает те или иные искажения и накладывает свои ограничения к применению на практике. В данной статье мы перечислим лишь некоторые модели.

- Стохастические модели (k-ограниченные стохастические модели, модели грамматики с конечным числом состояний и др.), предлагающие анализ на основе вероятностного распределения.
- Модели непосредственно составляющих (модель грамматических обязательств В. Ингве, модель непосредственных составляющих Ч. Осгуда, модели ядерных утверждений Ю.С. Степанова и Н.Д. Арутюновой, модель контекстуальной генерализации и др.), анализ которых производится путем разбора предложения на терминальные цепочки (или на ядерные утверждения, или на пропозиции).
- Модели трансформационной грамматики, опирающиеся на подход Н. Хомского (теория семантического компонента Дж. Каца и Дж. Фодора, концепция дифференциальных фонетических признаков Якобсона-Халле, и др.).
- Самое широкое прикладное распространение в методах анализа текста получили когнитивные модели (модель понимания речи В. Кинча, модель абстрактной грамматики языковой активности Ч. Осгуда, модель И. Шлезингера и многие другие), благодаря чему развилось отдельное направление – когнитивная лингвистика. Подробнее с данными когнитивными моделями можно ознакомиться в работе Б.М. Величковского [4].
- Существует еще целый ряд классов теоретических моделей, послуживших теоретической основой для конкретных методов анализа текста, как, например, психолингвистическая теория порождения речи (А.Р. Лурия, Н.И. Жинкин, А.А. Леонтьев) и множество других моделей, чьи постулаты позволяют диагностировать только узкий спектр свойств текста (например, фоносемантику).

Интересно заметить, что данные подходы направлены на реконструкцию из текста различных аспектов коммуникативной ситуации, основываясь на выделении и оценке соотношения тех или иных единиц анализа. Причем с течением времени характерно расширение контекста рассмотрения текста. Так психолингвистика первого поколения рассматривала текст, как «вещь в себе», вариации рассмотрения которого возможны лишь по внутренним законам формирования текста, а роль субъекта предопределялась врожденными и приобретенными правилами речепорождения. По мере развития психолингвистика расширяла представления, как о единицах анализа текста, так и о его возможностях данного метода (см. рисунок).

Таким образом, основываясь на вышеприведенной оценке методов анализа текста, мы можем выделить три критерия для систематизации всех методик и методов анализа текста: единица анализа, объяснительная модель анализа или сфера предназначения.

Систематизация методов анализа текста

Среди возможных решений систематизации множества разнородных методик анализа текста наиболее известны деления по выполняемым функциям и ориентировке.

А. По выполняемым функциям различают группу методов [14]:

а) способных осуществлять импорт текста и работу с ним;

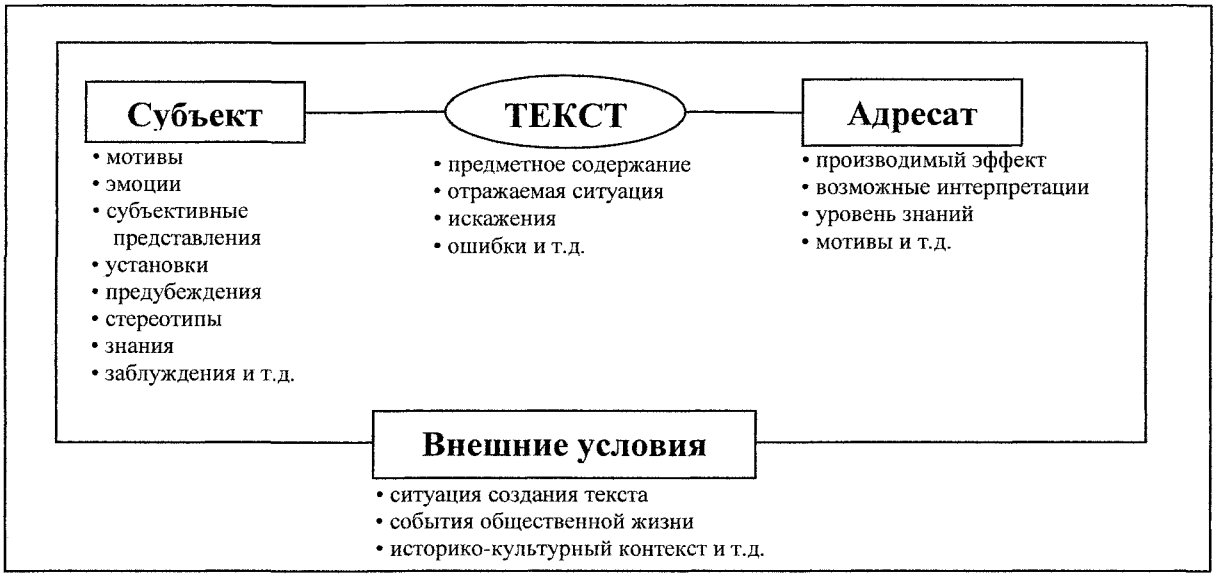
б) исследования текста (работают на грамматическом, синтаксическом уровне, осуществляют разнообразный поиск в тексте, выделяют ключевые слова, индексы и пр.);

в) ориентированных на семантический анализ, создание схем категоризации, словарей, кодирования;

г) позволяющих осуществлять экспорт данных анализа (например, сам текст или схему кодирования, или используемый словарь и т.п.).

В. По объекту анализа [21]

а) методы, «ориентированные на язык» (лингвистических единиц):



Сферы диагностики методов анализа текста

- лингвистические методы;
- методы работы с данными (поиск информации, списки слов, конкорданс, индексы и пр.).

б) методы, «ориентированные на контент» или содержательный анализ:

- *качественные методы*, позволяющие осуществлять поиск закономерностей и различий в тексте, анализировать текст целиком (некоторые методики позволяют анализировать аудио- и видеоинформацию). В данной группе методов для проведения качественного (содержательного) исследования текста могут быть использованы и количественные данные, которые помогают организовать качественную (содержательную) информацию. Важным отличием является преимущественное использование в качестве единиц анализа тем, концептов, процессов, контекстов. При этом объем анализируемого текста может быть ограничен;
- *методики анализа событий по текстовым данным;*
- *количественные методики*, позволяющие осуществлять статистическую проверку гипотез, ориентированы на исследование больших объемов текста:

– *категориальные* системы имеют встроенные или пользовательские словари, на основе которых осуществляется поиск в тексте (категории при этом могут быть как тематическими, так и семантическими), некоторые методики имеют ограничения по размеру единиц анализа;

– *некатегориальные* системы на основе одновременной встречаемости слов, строк, концептов позволяют строить разнообразные графы и дендрограммы;

– *системы кодирования ответов на вопросы незаконченных предложений* не предусмотрены для анализа больших объемов текста, предназначены для анализа достаточно гомогенного текста и лимитированы по размеру единиц анализа текста.

Работа по систематизации существующих методов крайне необходима для объединения различных способов содержательного анализа текста, так как несколько методов могут находить свою реализацию в одной конкретной методике, где данные по ним дополняют друг друга увеличивая репрезентативность и валидность результатов.

При этом под методикой анализа текста понимается «конкретной вариант того или иного метода, направленный на решение определенного класса исследовательских задач» [8, с. 73].

Ниже мы приводим названия соответствующих программ, по которым нетрудно найти сайты в Интернете. Часть из этих программ распространяется бесплатно, и заинтересованный читатель может установить их на своем компьютере и проверить в действии, другие имеют демоверсии, которые хоть и ограничены, но вполне адекватно позволяют представить возможности программы.

Интент-анализ – метод, позволяющий реконструировать интенции (субъективная

направленность на некий объект) автора по его тексту [12], поскольку для выявления и квалификации интенций опора на отдельные слова и предложения малопродуктивна. Экспертное выявление и идентификация речевых интенций предоставляет возможность очертить их круг в текстах разной тематики и направленности, т.е. охарактеризовать их качественно, поэтому исследовательская задача у использующих метод интент-анализа состоит в экспертном (т. е. по сути субъективном) оценивании характера интенций, их размытости и неясности понимания. Метод состоит из последовательных этапов: выделение круга обсуждаемых тем и вопросов, определение связей между объектами, затем кодификации дескрипторов, далее проводится оценка групп объектов по нескольким интегральным измерениям, полученные значения усредняются и определяются интегральные значения каждого объекта по указанным параметрам. При исследовании текстов СМИ интент-анализ позволяет решать проблемы социально-психологического и общесоциального плана – например, влияние средств массовой коммуникации на индивидуальное и групповое сознание. Как правило, вопрос о надежности получаемых при проведении интент-анализа результатов может быть решен путем введения контрольной процедуры, для которой привлекается дополнительная группа респондентов, которая высказывает согласие либо несогласие с квалификацией экспертов, проводивших первоначальную идентификацию. Поскольку метод интент-анализа основан на экспертной оценке текста, его полная автоматизация не осуществлена, однако, существуют методики, где автоматизированы первые этапы интент-анализа, позволяющие осуществлять операции выделения тем, кодирования и поиска. Отметим, к примеру, следующие.

- Ethnograph – методика предназначена для качественного (содержательного) анализа данных интервью, фокус-групп, дневников и пр. Процесс работы с данными построен на основе того, что производится поиск необходимых данных, отбор результатов поиска и анализ всего отобранного материала. Для каждого из звеньев этого процесса предусмотрен соответствующий набор процедур.
- Leximancer – методика, направленная на выявление ключевых тем (key-themes), концептов (concepts) в электронных документах. В отличие от других аналогичных методик, в своей основе соединяет множество подходов, таких как вычислительная лингвистика, контент-анализ, информационные науки, физика, теория сетей и пр. Позволяет наглядно представлять результаты анализа и вычислений в виде карты концептов и разнообразных таблиц и диаграмм.
- Minnesota Contextual Content Analysis (MCCA) – позволяет осуществлять контекстуальный анализ (анализ слов в пространстве 4 социальных контекстов: практика, традиции, эмоции, анализ) и анализ основных мыслей и образов, раскрытых в тексте. Для каждого вида анализа разработаны и стандартизированы нормы. Помимо указанных главных функций позволяет выводить статистику, производить анализ слов, частотный анализ слов и категорий и т.д.

Контент-анализ – самый распространенный метод, имеющий множество вариаций в различных методиках, позволяющий провести качественно-количественный анализ содержания текстовых массивов с целью последующей интерпретации выявленных числовых закономерностей. Заключается в оценке частотного распределения слов, словосочетаний словоформ и других единиц анализа (число их вариаций теоретически безгранично) относительно текста. Результатом является частота, относительный и удельный вес, вероятность встречаемости и пр., на основе чего делается качественный или количественный вывод в зависимости от выдвинутой гипотезы. Контент-анализ может быть проведен при помощи широкого спектра методик.

- **Crawdad Desktop** – методика, позволяющая осуществлять контент-анализ и поддерживающая сложное моделирование. В основу положена запатентованная технология. В отличие от других аналогичных методик, в которых частота встречаемости слова признается эквивалентной его значимости, данная методика базируется на лингвистической теории порождения людьми последовательностей речевых единиц в их коммуникации, благодаря которой создается сетевая модель текста. При этом значимость слова рассчитывается на основе структурной позиции слова в полученной сети.
- INTEXT – методика, обладающая разнообразным функционалом, позволяющая создавать списки слов, разнообразные последовательности слов, перестановки слов, перекрестные ссылки, производить сравнение

списков слов, контент-анализ, составлять конкорданс, управлять данными и др.

• Kwalitan – методика, позволяющая осуществлять качественный (содержательный) анализ таких данных, как протоколы интервью и наблюдений, разнообразные тексты (статьи, античные рукописи и пр.). Позволяет осуществлять кодирование, поиск, анализ слов и кодов, подготовку материалов для дальнейших исследований.

• PROTAN – автоматизированная система контент-анализа, которая предназначена для анализа любого текста (рассказы, клинические интервью, научные публикации, названия или резюме научных журналов, поэзия, рекламные материалы и др.). Ограничения PROTAN обусловлены статистическими ограничениями, отсутствием словарей, необходимых для того, чтобы анализировать специфический вид текста. Текст для анализа должен быть представлен в стандартной кодировке. Методика позволяет решать две основные задачи: определение структуры текста (при помощи семантических словарей), определение основных тем и идей текста (на основе информации, содержащейся во взаимосвязях между единицами анализа текста).

• Yoshikoder – методика контент-анализа текста, включающая разработку и использование словарей, автоматический поиск по ключевым словам в контексте.

Фоносемантический анализ текста или слова заключается в оценке его звучания безотносительно к его содержанию. Заключается в сопоставлении системы сочетаний фонем в конкретном тексте или слове с их стандартизированными оценками по ряду биполярных шкал [6]. Результатом фоносемантического анализа является профиль выраженности оценочных шкал в стандартизированном семантическом пространстве, на основании которого делается заключение о возможном воздействующем эффекте текста на читателя. Однако ввиду крайней специфичности единиц анализа и непроработанного механизма соотнесения с содержанием текста (и отсутствия контроля факторов, влияющих на процесс осмысления текста) результаты этого метода представляются многим исследователям сомнительными и не обладающими внешней валидностью, что не отменяет адекватность применения данного метода для узкоспециализированных исследований. Фоносемантический анализ реализован в следующих методиках.

• Vaal – в методике реализованы алгоритмы оценки фонетического воздействия на человека слов и текстов русского языка, причем в основе этого эмоционального воздействия фоники слова и текста на подсознание человека лежат психофизиологические механизмы. Возможно анализировать готовые тексты с точки зрения такого воздействия, составлять новые с заданным вектором воздействия, выявлять личностно-психологические качества авторов текста, проводить углубленный контент-анализ и делать многое другое.

• DIATON – это методика экспертизы суггестивных текстов, основанная на фоносемантическом анализе и ориентированная на оценку скрытых особенностей, которые сложно осознать: фоносемантических, ритмических, структурных характеристик текста.

Дискурс-анализ или дискурсивный анализ – совокупность методик и техник интерпретации текстов или высказываний как продуктов речевой деятельности, осуществляемой в конкретных общественно-политических обстоятельствах и культурно-исторических условиях [17]. Этот метод ориентирован, прежде всего, на изучение лингвистического уровня в структуре социальной коммуникации как доминирующего на протяжении определенного исторического периода развития общества и культуры. Сам метод заключается в последовательности ряда операций: фиксация изучаемого материала; выделение его формальных характеристик; обозначение контекста как коммуникативной ситуации; выбор направления и стратегии анализа; теоретическое дифференцирование и структурирование этапов исследования; определение техники и средств анализа при использовании конкретной модели исследования; дефиниции единиц анализа; проверка системы категорий в теории и на эмпирическом материале; осуществление основных этапов исследования (дескрипция, реконструкция, интерпретация); фиксация результатов исследования, их обобщение, истолкование и структурирование. Дискурс-анализ позволяет выделить не только существенные характеристики социальной коммуникации, но и второстепенные, содержательные и формальные показатели (например, тенденции в вариативности речевых формул или построении высказываний). Дискурс-анализ широко применяется в социологических и политических исследованиях и отчасти реализован в таких программах, как CATPAC. Это методика анализа текста, напи-

санного на любом языке, основанная на системе Galileo, которая представляет собой комплекс теории и методов, направленных на научное изучение когнитивных и культурных процессов. CATPAC позволяет выявлять основные идеи текста без предварительного кодирования и лингвистического анализа.

Нарративный анализ – это метод обобщения прошлого опыта при помощи соотнесения последовательности слов в предложении и последовательности реальных (как предполагается) событий [22]. Позволяет осуществлять количественную оценку текста. В отличие от контент-анализа, который может быть применен к любым текстам, нарративный анализ ориентирован на особые тексты, содержащие рассказ. Преимуществом нарративного анализа, по сравнению с кластерным, является то, что оценка производится по конкретным категориям (Субъект, Действие, Объект), а не по произвольно выбранным исследователем исходя из его задач. В класс нарративных текстов входят разнообразные истории от разнообразных художественных и исторических текстов (мифы, легенды, летописи и пр.) до статей газет, в которых описываются произошедшие события. Нарративный анализ используется совместно с другими методами анализа текста и реализован в известных методиках.

- LIWC – методика, обладающая 68 встроенными словарями (лингвистические, психологические конструкты и др.), представляющими собой пространство для оценки того, в какой степени испытуемые используют слова тех или иных категорий (например, позитивные и негативные слова). Методика позволяет осуществлять нарративный анализ текста. Данная методика также позволяет осуществлять синтаксический анализ, интент-анализ и ряд других функций.
- PC-ACE – методика, предназначенная для кодирования событий, позволяющая организовывать сложную текстовую информацию, хранить ее, осуществлять поиск необходимых данных, характеризующихся сложной структурой. Методика применима для всего спектра социальных наук. Позволяет осуществлять качественный (содержательный) анализ данных [19].

Экспертная оценка текста – в эту группу методов входят различные экспертизы текста, классификации которых, согласно А.А. Леонтьеву [11], можно представить в следующем виде: а) авторoведческая экспертиза, направленная на установление автора текста или выявление категориальных признаков вероятного автора: пол, возраст, национальность, место рождения, место долговременного проживания, уровень образования и пр. [3]; б) экспертиза, направленная на установление временных признаков автора текста (эмоциональное состояние и пр.); в) экспертиза, направленная на установление тех или иных условий создания исследуемого текста (также экспертиза аутентичности записей при интервью); г) экспертиза, направленная на установление преднамеренного искажения сведений, высказываемых в тексте; д) экспертиза, направленная на установление определенных признаков (оскорбление, призыв и пр.). Для осуществления данных экспертных оценок применяется комплекс методик: а) методика перефразирования текста или законченного фрагмента текста, б) методики семантического шкалирования, например, методика семантического интеграла (Батов В.И., Сорокин Ю.А.); в) методики свободного ассоциативного эксперимента; г) методики предикативного анализа текста. Существуют также компьютеризированные варианты данных экспертиз [2].

Графематический анализ – метод, создающий базу для последующего морфологического и синтаксического анализа на основе выделения слов, цифровых комплексов, формул и т. д. Анализ направлен на разбивку текста на слова, разделители и т. д.; сборку слов, написанных в разрядку; выделение устойчивых оборотов, фамилии, имени, отчества, даты и т. п; выделение электронных адресов и имен файлов; выделение предложений из входного текста абзацев, заголовков, примечаний.

Морфологический анализ направлен на определение множества морфологических интерпретаций каждого из слов текста, состоящей из таких параметров, как лемма, морфологическая часть речи; набор общих граммем; множество наборов граммем. Морфологический анализ реализован в большинстве методик, так как является основой для других видов анализа текста. В качестве примеров можно отметить реализацию данного метода.

- ATLAS.ti – методика, позволяющая анализировать большие объемы текста, разнообразные графики, аудио- и видеоинформацию. Может применяться в социальных и экономических науках, маркетинге и менеджменте, теологии. Методика рассчитана на проведение качественного (содержательного) анализа данных и включает средства ис-

следования текста, управления текстом, сравнения и пр.

- Textanz – методика, предназначенная для частотного анализа текста на уровне слов, фраз, словоформ. С помощью данной методики возможно производить анализ любого текста.
- TextArc – методика, позволяющая наглядно представлять и просматривать текст. С ее помощью можно подсчитывать разнообразные индексы, составлять конкорданс, автореферат. Также данная методика применима для синтаксического анализа.

Синтаксический анализ – это метод сопоставления линейной последовательности лексем языка с его формальной грамматикой. Результатом анализа становится синтаксическая структура предложения, которая представляется в виде дерева зависимостей. Результаты синтаксического анализа важны для дальнейших этапов работы с текстом. Синтаксический анализ текста позволяют провести следующие методики.

- Profiler Plus – многоцелевая методика, предназначенная для кодирования текста. Включает синтаксический анализ и некоторые другие, более сложные принципы кодирования, использующие обобщения, множественные переходы и пр. Ориентирован на использование в сфере бизнеса, политике и в академической среде. Текст для анализа может быть на английском, русском, испанском, арабском или китайском языках. Полученные по итогам анализа результаты могут быть обработаны в статистических пакетах (например, SPSS), для оценки полученных результатов применяется экспертная оценка.
- DictaScope – методика для проведения синтаксического анализа. Производится анализ подчинительных связей между словами на основе принадлежности этих слов к той или иной части речи.

Семантический анализ – метод, направленный на построение семантической структуры предложения, состоящей из семантических узлов и семантических отношений. Целью проведения анализа является построение этих узлов, которые образуются из слов исходного предложения. Основу для формулирования гипотез относительно состава семантических узлов составляет информация, полученная в результате синтаксического анализа. Результаты анализа представляются в виде семантического графа, построение которого состоит из ряда этапов (инициализация семантических узлов и синтаксических вариантов фрагментов, построение множества словарных интерпретации узлов, построение групп времени, построение узлов в кавычках и т. д.). Семантический анализ может быть осуществлен при помощи различных методик, например, PROTAN (рассмотренной нами выше) и широком спектре иных методик. Так, например, семантический анализ реализован в методике T-LAB Tools for Text Analysis, которая представляет собой компьютерную методику, позволяющую осуществлять три вида анализа: тематический анализ, сравнительный анализ, анализ смежности, а также выявляет смысловые паттерны слов и основных идей текста. Процесс работы с текстом в этой методике включает в себя сегментирование текста, отбор ключевых слов, а также процедуры, предназначенные для осуществления трех типов анализа.

Кроме вышеперечисленных методов существует еще целый ряд методов анализа текста: структурный анализ, семиотический (семиологический) анализ, системный анализ, символический (мифологический) анализ; анализ социальных индикаторов и наррации (линии) ключевых слов; социально-ролевой анализ, риторический анализ, перформативный анализ, жанровый анализ, речедеятельностный анализ, психоаналитический анализ, критический анализ, исторический анализ, культурологический анализ, интертекстуальный анализ, феноменологические типы анализа; анализ коммуникативных стратегий и свободных ассоциаций; прагма-, психо-, социо-, этно-, когнитивно-лингвистический анализ и т. д. Неоспоримым преимуществом во всем разнообразии методов анализа текста обладает контент-анализ, модификации и разновидности которого позволяют решать самые разнообразные исследовательские задачи [15].

Важными параметрами при проведении любого анализа текста выступают достоверность полученных данных (обеспечиваемая полнотой анализируемого текста и его репрезентативностью) и надежность интеркодирования единиц анализа, полностью зависимая от квалификации исследователя и теоретической модели, лежащей в основе. Ограничения применения методов анализа текста связаны, прежде всего, с субъективным влиянием исследователя, определяющего выбор единиц анализа и интерпретацию полученных числовых данных. Подробнее о проблемных зонах метода [9].

Вопросы валидности кодировки в методах анализа текста

Как уже упоминалось выше, главной основой в методах анализа текста выступает выбранная единица анализа. Выбор кодировки, как правило, осуществляется исследователем или группой экспертов, которые определяют единицу анализа и форму кодировки в зависимости от целей и гипотез проводимого исследования. От правильности и адекватности этого выбора зависит как валидность проводимого в дальнейшем статистического анализа, так и полученные результаты. Уникальность текста как продукта деятельности автора, возможность различной трактовки текстовых фрагментов делает кодирование информации в этом тексте необходимой процедурой (как в статистических, так и в качественных исследованиях), при которой неизбежно происходит обобщение, искажение и влияние личных проекций исследователя. Опыт использования методов анализа текста показал следующие пути для повышения валидности кодировки.

1. Создание особых словарей-таксономий (словарей речевых форм, понятий), в которых за каждым словом приводится описание той психической составляющей, о выраженности которой эта речевая форма или слово может свидетельствовать. Эти словари, как правило, выглядят в форме дендрограммы и состоят из слов, субкатегорий, категорий, суперкатегорий и т.д. К таким словарям относятся, например, таблица речевых форм Р. Хогенраада [20], построенная на теории Мак Клелланда [23], или словарь М. Брэдли и П. Ланга аффективных норм английский выражений [16]. Основной проблемой здесь является проведение четких разграничений между категориями и проведение исследований по всему спектру используемых в словарях категорий. Важным также является то, что такие словари довольно быстро могут устаревать, так как за употреблением конкретного слова у среднего респондента может меняться его смысловое содержание и, следовательно, психическое содержание которое стоит за употреблением этого слова [18].

2. Согласованный выбор единицы анализа и кодировки несколькими экспертами также может выступать в качестве фактора, увеличивающего валидность кодировки: когда несколько экспертов оценивают текст схожим образом. Повышение требований при этом к самому кодировщику оправдано особенно в тех случаях, когда выборка респондентов мала или сами тексты имеют небольшой объем – это связано с тем, что влияние допущенной ошибки или искажения при кодировке может привести к неадекватному результату всего исследования [1]. Особенно распространено искажение, допускаемое при кодировке, когда исследователь сознательно «укрупняет» некоторые категории кода в ущерб другим для подтверждения выдвинутой гипотезы.

3. Вторичная кодировка другим кодировщиком того же текста может быть использована как раз в тех случаях, когда полученные результаты исследования текста выглядят противоречиво, а сам текст малоинформативен и существует вероятность неправильной кодировки неоднозначных выражений. Сопоставление результатов может показать устойчивые закономерности в тексте или наоборот выявить недостатки кодирования.

4. Использование обратной связи от авторов текста в ситуациях, когда получен непредвиденный или необъяснимый результат, может оказаться полезным не только при интерпретации полученных данных, но и для проверки правильности выполненной кодировки, поскольку восприятие текста кодировщиком может кардинально отличаться от авторского в случае сложного по стилистике текста или текста, содержащего стилистические ошибки и смысловые искажения. Также это бывает полезно при малых объемах анализируемого текста или стенограммы речевого выступления.

5. В некоторых случаях для оценки выраженности психологических составляющих автора, стоящих за его текстом, используется метод сопоставления характеристик текста (семантических, частотных, морфологических и пр.) со стандартизированными на большой выборке нормами [13], где отклонение от норм рассматривается как свидетельство о выраженности психологических особенностей автора. Однако такой анализ, несмотря на то, что он дает интересные и информативные результаты, может выступать лишь в качестве косвенного способа проверки гипотезы, поскольку отличительные от нормы особенности текста не всегда являются следствием именно психологической составляющих, а могут быть обусловлены ситуативно.

6. Для контроля за ситуативным влиянием на исследуемые переменные хорошо подходит мониторинг цикла текстов автора, где можно определить стабильные во времени и от ситуации выраженные психологические характеристики автора.

Как можно увидеть из перечисленных способов повышения валидности анализа текста, создание совершенных компьютеризированных методик оценки текста будет сталкиваться с многочисленными проблемами при их применении. Однако увеличивающаяся доступность и объем текстовой информации делает анализ текста привлекательным способом оценки психологических особенностей автора или адресата информации. В этих условиях нужно помнить о способах применения этих методов в психологических исследованиях, поскольку анализ продуктов деятельности, по сути, является косвенным качественным методом и требует сочетания с количественными данными и данными прямых методов диагностики для получения достоверных и репрезентативных результатов.

Выводы

Приведенные примеры подходов далеко не исчерпывают всей сферы исследования текстов. Расширение контекста методов применения анализа текста, а также поиск зависимостей и связей единиц анализа текста с психическими и социальными процессами создало большое разнообразие подходов и способов исследования. Нами были представлены два критерия их систематизации: по выполняемым функциям и по объекту анализа. При этом было отмечено, что отправным моментом, который объединяет все методики анализа текста, является то, что в их основе лежат представления о единице анализа. Понятие единиц анализа является крайне важным аспектом, поскольку выступает своего рода аналогом исследуемых (но неконтролируемых) переменных в эксперименте. Нами были показаны различия лингвистических, психолингвистических, психологических единиц анализа текста и приведено систематизированное описание разработанных методик анализа текста, имеющих программную реализацию. Существующее разнообразие методик анализа текста мы постарались систематизировать в 10 групп: интент-анализ (Ethnograph, Leximancer, Minnesota Contextual Content Analysis), контент-анализ (Crawdad Desktop, INTEXT, Kwalitan, PROTAN, Yoshikoder), фоносемантический анализ (Vaal, DIATON), дискурс-анализ (CATPAC), нарративный анализ (LIWC, PC-ACE), экспертная оценка текста, графематический анализ, морфологический анализ (ATLAS.ti, Textanz, TextArc), синтаксический анализ (Profiler Plus, DictaScope), семантический анализ (PROTAN, T-LAB Tools for Text Analysis).

Литература

1. Алмаев, Н.А. Валидность кодировки контент-анализа и новые возможности ее оперативной оценки и корректировки (на материале интервью пациента с игровой зависимостью) / Н.А. Алмаев, В.И. Олешкевич // Проблемы психологии дискурса / под ред. Н.Д. Павлова, И.А. Зачесова. – М.: Изд-во «Институт психологии РАН», 2005. – С. 73–84.

2. Батов, В.И. Алгоритмизация некоторых процедур автороведческой экспертизы / В.И. Батов // Актуальные проблемы теории и практики применения математических методов и ЭВМ в деятельности органов юстиции. – М., 1975 – Вып. 4. – С. 85–88.

3. Батов, В.И. Опыт построения методики для установления авторства текстов / В.И. Батов, Ю.А. Сорокин // Известия АН СССР, Сер. литературы и языка. – 1977. – 36(4) – С. 345–347.

4. Величковский, Б.М. Современная когнитивная психология / Б.М. Величковский. – М.: Изд-во Моск. ун-та, 1982. – 336 с.

5. Выготский, Л.С. Избранные психологические исследования / Л.С. Выготский. – М.: Изд-во АПН РСФСР, 1956. – 519 с.

6. Журавлев, А.П. Фонетическое значение / А.П. Журавлев. – Л.: Изд-во Ленингр. ун-та, 1974. – 150 с.

7. Зимняя, И.А. Лингвопсихология речевой деятельности / И.А. Зимняя. – М.: Воронеж, 2001. – 432 с.

8. Леонтьев, А.А. Основы психолингвистики / А.А. Леонтьев. – М.: Смысл, 1997. – 287 с.

9. Мангейм, Дж. Б. Политология: Методы исследования / Дж. Б. Мангейм, Р.К. Рич. – М.: Изд-во «Весь Мир», 1997. – 544 с.

10. Николаева, Т.М. Единицы языка и теория текста / Т.М. Николаева // Исследования по структуре текста. – М., 1997. – С. 27–57.

11. Психолингвистическая экспертиза ксенофобии в средствах массовой информации: методические рекомендации для работников правоохранительных органов. – М.: Смысл, 2003. – 85 с.

12. Слово в действии. Интент-анализ политического дискурса / под ред. Т.Н. Ушаковой, Н.Д. Павловой. – СПб.: Алетейя, 2000. – 316 с.

13. Шалак, В.И. Контент-анализ: приложения в области политологии, психологии, социологии, культурологии, экономики и рекламы / В.И. Шалак. – М., 2004. – 272 с.

14. Alexa, M. Text Analysis Software: Commonalities, Differences and Limitations: The Results of a Review / M. Alexa, C. Zuell // Springer Netherlands, – 2000. – 34 (3) – P. 299-321.

15. Berelson, B. Content Analyses in Communication Research / B. Berelson. – Glencoe, 1952. – 220 p.

16. Bradley, M. Affective norms for english words (anew): Stimuli, instruction manual and affective ratings. / M. Bradley, P. Lang // Technical report c-1. – Gainesville, FL: University of Florida, 1999. – P. 1–49.

17. Brown, G. Discourse analysis / G. Brown, G. Yule. – Cambridge: Cambridge University Press, 1983. – 288 p.

18. Dodds, P. Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents / P. Dodds, Ch. Danforth //Journal of Happiness Studies. – 2009. – 11(4). – P. 441–456.

19. Franzosi, R. Quantitative Narrative Analysis (Quantitative Applications in the Social Sciences) / R. Franzosi. – Beverly Hills, CA: Sage, 2009. – 200 p.

20. Hogenraad, R. Force and influence in content analysis: The production of new social knowledge / R. Hogenraad, D.P. McKenzie, N. Puladeau // Quality & Quantity. – 2003. – 37(1). – P. 221–238.

21. Klein, H. Classification of Text Analysis Software / H. Klein // Classification and knowledge organization: Proceedings of the 20th annual conference of the Gesellschaft fur Klassifikation / In R. Klar, O. Opitz (Eds.). – University of Freiburg, 1996, Berlin, New York: Springer, 1997. – P. 255–261.

22. Labov, W. Sociolinguistic patterns / W. Labov. – Pennsylvania: University of Pennsylvania Press, 1972. – 346 p.

23. McClelland, D.C. Power: The Inner Experience / D.C. McClelland. – New York: Irvington Publishers, 1975. – 427 p.

24. Saporta, S. Relation between Psychological and Linguistic Units Psycholinguistics / S. Saporta. – Baltimore, 1954. – 325 p.

Поступила в редакцию 14 октября 2010 г.

Митина Ольга Валентиновна. Канд. психол. наук, вед. науч. сотр. факультета психологии, Московский государственный университет им. М.В. Ломоносова: omitina@inbox.ru.

Olga V. Mitina. Ph.D. Sci. Sciences, Senior Research Officer Department of Psychology at Moscow State University named after M.V. Lomonosov Moscow State University: omitina@inbox.ru.

Евдокименко Александр Сергеевич. Канд. психол. наук, науч. сотр. факультета психологии, Московский государственный университет им. М.В. Ломоносова: belbel@list.ru.

Alexander S. Evdokimenko. Ph.D. Sci., Research Fellow Department of Psychology at Moscow State University named after M.V. Lomonosov Moscow State University: belbel@list.ru.