УДК 808.2:159.937

Невдах М. М., аспирант

ПРИМЕНЕНИЕ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ В ИССЛЕДОВАНИИ ТЕКСТОВ

In the article application various quantitative methods in studying the text as statistical set is considered, and also the domestic on the basis of given methods domestic and foreign software products intended for the analysis and linguistic processing of texts are described. A number most the pressing questions demanding more detailed studying is allocated: it, first of all, researches in readability area with use of modern information technologies and development of corresponding toolkit for classification of Russian-speaking texts on a number of fields of knowledge depending on readiness of the reader.

Введение. С возникновением и развитием ряда наук, в которых центральным объектом анализа выступает текст, подтвердилось предположение о том, что текст представляет собой структуру, элементы которой подчиняются законам, определяющим статистическую упорядоченность и строгую организацию. Точный характер проявляющихся закономерностей, регулярностей в языке в целом крайне сложно уловить без применения математических методов и ЭВМ. Поэтому интерес к квантитативным методам как инструменту научного и практического познания статистических свойств языковых структур повышается и обусловлен объективной реальностью.

Основная часть. Текст как статистическая совокупность может быть охарактеризован через множество количественных переменных, на основе которых он преобразуется из последовательности символов в набор чисел. Особенностью этих переменных является то, что они по определению не отражают глубинных, сущностных сторон текста, а описывают только внешнюю, поверхностную сторону текста. При этом многие исследователи полагают, что формальные признаки каким-то опосредованным, вероятностным образом связаны с содержательной сущностью текста. В связи с этим набор количественных признаков часто является диагностическим при решении конкретной задачи (например, атрибуции текста, оценке его трудности), что, несомненно, открывает путь для проникновения в глубинную организацию текста, не доступную непосредственному наблюдению.

Замена словесного описания текста его математическим представлением в компьютерной среде позволяет избежать бесконечного богатства ассоциаций, возникающего при «живом» общении с текстом, и при этом вскрыть характер закономерностей, присущих определенным языковым структурам.

Квантитативный анализ текстов в настоящее время позволяет решать различные научно-практические задачи. В работах, связанных с изучением текстов как статистических объектов, преобладают исследования, направленные на оценку близости и однородности стилей текстов и их классификацию.

Задача проверки близости стилей состоит в том, чтобы сравнить два или более текстов, заданных совокупностью количественных признаков, и установить различие между стилями. После попарного или множественного сравнения стилей есть возможность установить различие в виде альтернативы «да/нет» либо в виде значения степени различия стилей.

Первая работа в этом направлении принадлежит Т. Менденхоллу [1], в которой автор сравнивает стили текстов произведений различных писателей, написанных как на одном языке, так и на разных. Сравнение проводится на основе гистограмм, которые отражают частоту появления слов с разной длиной.

Похожий подход для сравнения стилей использовал Н. А. Морозов [2]. В качестве признаков стиля исследователь использовал частоту появления наиболее встречающихся слов (предлогов, союзов, частиц). На основании гистограмм распределений данных слов Н. А. Морозов проверял близость стилей.

Среди работ, написанных в последние годы, следует отметить диссертацию О. Г. Шевелева [3], в которой разработаны алгоритмы и инструментарий для сравнения стилей текстовых произведений. В частности, предложены новые подходы для сравнения стилей текстов с использованием гипергеометрического критерия (двустороннего точного критерия Фишера) и критерия хи-квадрат по отдельным частотным признакам текстов, совокупности признаков, а также по их распределению; предложен новый подход к кластеризации текстов с использованием таких мер сходства, как «частота рассогласования» и интегральная мера рассогласования; предложены модификации метода Хмелева классификации текстов по авторскому стилю с использованием для оценки расхождения частот мер Кульбака и хи-квадрат. Автором также создан программный комплекс «СтилеАнализатор» для сравнения стилей текстов.

Смежной по отношению к задаче проверки близости стилей текстов является задача проверки текстов на однородность стиля. Методы проверки текстов на однородность могут использоваться для сравнения стилей, и наоборот, методы сравнения стилей — для проверки однородности текстов.

Наиболее известным методом проверки текстов на однородность авторского стиля является метод накопительных сумм [4], суть которого заключается в том, чтобы выбрать несколько характеристик, являющихся функциями предложения. Например, для английского языка А. К. Мортон использовал длину предложения и число двух- и трехбуквенных слов плюс число слов, начинающихся с гласной буквы. После этого производился расчет этих характеристик для каждого предложения, вычислялись их средние значения. По отклонениям от средних значений для каждого предложения строилась накопительная сумма. Для однородного стиля графики характеристик практически совпадали.

Для проверки однородности текста используется и метод структурного анализа текста, предложенный отечественными исследователями А. Ф. Толочко и Н. И. Миницким [5]. Разработанная авторами математическая модель на основе изучения учебного текста позволяет создать функцию возмущений частоты появлений отдельных букв алфавита на заданной выборке относительно этих же характеристик на генеральной совокупности. В качестве эталона-образца может использоваться норма русского языка или других языков, характеристика стиля автора на генеральной совокупности текста учебника, который признан в педагогическом сообществе образцовым.

Н. С. Закревская в [6] рассматривает подход к проверке однородности, основанный на проверке соответствия числовых последовательностей модели фрактального броуновского движения. Числовые последовательности получены путем замены слов текста на их длины, измеренные в слогах.

Набор алгоритмов, позволяющих производить классификацию и идентификацию изучаемых объектов, в научной литературе принято обозначать термином «распознавание образов». При этом задачей классификации является построение алгоритма классификации, т. е. правила отнесения предъявляемого объекта к тому или иному классу.

Когда в качестве объектов выступают тексты, то наиболее часто исследователи решают задачу классификации текстов по авторству. Среди методов, связанных с атрибуцией, можно выделить энтропийный метод Д. В. Хмелева [7] и метод О. Хрулева, основанный на использовании частотного словаря [8].

Метод Д. В. Хмелева позволяет с высоким качеством (84%) классифицировать тексты по авторству на основе формальной математической модели последовательности букв текста как реализации цепи А. А. Маркова. Для выбранных текстов вычисляется матрица переходных частот употребления пар букв. Она служит оценкой матрицы вероятностей перехода из буквы в букву. Автором анонимного текста полагается тот, у которого вычисленная оценка вероятности больше. Существуют и другие исследования Д. В. Хмелева, в которых при разработке методики определения авторства учитываются такие формальные характеристики языка автора, как число служебных слов (предлогов, союзов и частиц), используемые морфемы (приставочные, корневые, суффиксальные, флективные) и их последовательности, сложность используемых грамматических конструкций и собственно словарь, используемый автором. Каждый из параметров использован в модели ЛингвоАнализатора, позволяющей определять наиболее вероятное авторство.

Метод О. Хрулева позволяет классифицировать тексты по авторству на основе сравнения частотных словарей писателей. В словарь входят 10 000 наиболее употребительных слов русского языка. Полученные частоты для каждого писателя делятся на средние частоты в русском языке, взятые из частотного словаря С. А. Шарова. Писатель определяется по наименьшему расстоянию между словарями писателей и словарем анализируемого текста. Расстояние определяется как сумма разностей частот между отдельными анализируемыми словами. Для текстов, участвовавших в формировании словарей, частота правильных классификаций составляет 98%.

Для классификации текстов используются и другие, более сложные методы: нейронные сети; метод опорных векторов; классический дискриминантный анализ; вероятностный классификатор; метод сжатия данных; методы, основанные на извлечении правил (методы накопительного извлечения правил, деревья решений, метод «колонии муравьев»).

Перечисленные подходы и методы позволяют в настоящее время решить ряд вопросов, связанных с систематизацией и изучением текста. Благодаря точным математическим методам открываются возможности для анализа скрытых потенциальных возможностей текста. Наработки в этой области можно с успехом применить во многих сферах, в том числе и редакционно-издательской деятельности. Во-первых, появляется возможность тестировать стили авторского коллектива (в случае, когда несколько авторов пишут одну книгу) на предмет их близости, однородности. Это особенно важно в сфере учебного книгоиздания. Во-вторых, можно проанализировать лингвостатистические характеристики текстов и дать рекомендации по их корректировке. И в-третьих, можно установить атрибуцию текста, что очень важно для текстологической науки. Кроме того, важной является информация и о том, использовал ли автор при написании произведения

дополнительные источники (например, сеть Интернет). Это может быть серьезным аргументом при экономических расчетах с автором.

Несмотря на разноплановые исследования в области квантитативного анализа текстов, один из важнейших вопросов остается недостаточно разработанным. Данная проблема связана с оценкой трудности текста для будущих читателей, решение которой будет являться важным шагом в повышении качества подготовки литературы, что имеет особое значение при выпуске учебных изданий.

В отечественной науке в настоящее время практически отсутствуют объективные инструменты для классификации текстов в зависимости от подготовленности читателей. В определенной степени вопросы количественного анализа текстов и выявления факторов, влияющих на усвоение материала, раскрыты в работах, связанных с читабельностью текста [9].

На данный момент существуют компьютерные программы, предназначенные для анализа и лингвистической обработки текстов. Однако следует отметить, что провести всестороннюю обработку текстов в рамках какой-то одной программы невозможно. Каждый программный продукт направлен на решение конкретных прикладных задач.

Одним из наиболее известных продуктов для классификации текстов по авторству является система «ЛингвоАнализатор» Д. В. Хмелева, доступная на сайте автора по адресу http://www.rusf.ru. Программа определяет возможного автора текста (выдает имена трех писателей) среди 128 писателей, заложенных в систему. Кроме того, ЛингвоАнализатор находит три произведения каждого из авторов, которые наиболее близки данному тексту. Применяемая методика определения авторства опирается на математическую модель, в которой учтены формальные характеристики языка автора. Набор авторов, их тексты и признаки авторских стилей для алгоритма (но не для обработки) заложены в программу. Возможность их изменения со стороны пользователя не предусмотрена.

Информационная система «СМАЛТ» (Статистические методы анализа литературного текста) [10], разработанная в Петрозаводском государственном университете, позволяет произвести настраиваемый анализ от выбора текстов до конечного представления результатов анализа. Блок-анализ состоит из трех основных модулей. Первый модуль ориентирован на выборки из базы данных, основанные на лингвостатистических параметрах (например, общее распределение длины слов и предложений, средняя длина предложения в словах, индекс разнообразия лексики и т. д.). Модуль допускает задание объема выборки, а также проверки статистических гипотез о равенстве средних на основе критерия Стьюдента и проверки данных на однородность при помощи непараметрического критерия Колмогорова-Смирнова. Второй модуль предназначен для реализации методики атрибуции, основанной на изучении закономерности расположения частей речи в рамках предложения. Третий модуль позволяет измерять близость текстов на основе методов кластерного анализа: иерархической кластеризации, метода корреляционных плеяд и т. д.

В автоматизированной системе обработки лингвостатистических данных «ЛИНДА» [11], разработанной на кафедре структурной, прикладной и математической лингвистики Санкт-Петербургского государственного университета, решаются следующие задачи:

а) первичная обработка лингвистических данных (построение рядов распределения, вычисление статистик, статистических оценок, проверка статистических гипотез и др.);

б) лексикографическая обработка текстовых данных: создания частотных и алфавитно-частотных словарей, словарей-конкордансов, словоуказателей, обратных словарей, словарей ключевых слов и т. п.;

в) информационно-поисковые, в том числе поиск текстовых единиц, обладающих определенным набором количественных и качественных характеристик для решения стилистических и грамматических проблем; автоматический поиск текстов (авторский, жанровый, историко-хронологический и др.);

г) систематико-таксономические, в том числе обработка многомерных данных с использованием стандартных алгоритмических процедур (кластерного, факторного и других методов многомерного анализа); обработка лингвистических данных с помощью специальных лингвистических методов (дешифровочных алгоритмов, методов датировки, атрибуции, диагностики и типологии текстов и др.);

д) теоретико-статистические исследования: изучение статистических закономерностей в символьных последовательностях, изучение проблем устойчивости и вариативности лингвостатистических чисел, проблемы однородности текстов, условий действия закона больших чисел, оптимизация выборочных исследований и др.

Одной из самых мощных систем аналитической обработки, позволяющей работать с текстами, является PolyAnalyst [12]. Основная функциональность программы предназначена для извлечения знаний из больших баз данных. В аналитический инструментарий системы входят модули для построения числовых моделей и прогноза числовых переменных, алгоритм кластеризации, алгоритмы классификации, алгоритмы ассоциации, модули визуализации данных.

Для работы с текстом в PolyAnalyst предусмотрен модуль TextAnalyst, являющийся средством формализации неструктурированных текстовых полей баз данных. В модуле предусмотрены построение семантической сети понятий, выделенных в обрабатываемом тексте, со ссылками на контекст; смысловой поиск фрагментов текста с учетом скрытых в тексте смысловых связей со словами запроса; анализ текста путем построения иерархического дерева тем/подтем, затрагиваемых в тексте; реферирование текста.

Система DICTUM (система для универсальной обработки и анализа словарей и текстов) разрабатывается и используется лабораторией общей и компьютерной лексикологии и лексикографии филологического факультета МГУ с 1991 г. Эта система позволяет создавать, расширять, сравнивать, объединять словари, осуществлять по ним сложный поиск, включающий грамматические, частотные и другие характеристики, делать привязку словарных статей к определенным местам какого-либо текста. Подсистема обработки текстов производит разметку текстов как признаками, заданными извне (например, название, жанр), так и извлеченными в процессе анализа его внутренней структуры. Подсистема позволяет производить лексический, морфологический и синтаксический анализ. Аналитические инструменты включают в себя морфолемматизатор, поиск повторяющихся фраз, инструмент для пополнения и использования семантических характеристик слов и фраз и некоторые другие. Среди баз данных DICTUM имеются базы синонимов, омонимов, идиом, морфем, грамматически размеченных слов, тезаурус.

Следует также отметить семейство программных продуктов, выпускаемых под торговой маркой RCO, которое предназначено для решения задач, требующих автоматического анализа текста на русском языке. Разработанное лингвистическое и алгоритмическое обеспечение позволяет решать такие прикладные задачи, как составление содержательного портрета текста, извлечение именованных объектов, связей и фактов из массивов неструктурированных данных, анализ тональности текста, выявление заимствований и дубликатов.

Экспертная система «ВААЛ» производит количественный анализ текстов, но для решения психолингвистических задач: прогноза эффекта неосознаваемого воздействия текста на массовую аудиторию, анализа текстов с точки зрения такого воздействия, генерации текста с заданным вектором воздействия, выявления личностно-психологических качеств автора текста. Система позволяет оценивать слова с точки зрения их фоносемантического воздействия на человека; задавать желаемые фоносемантические характеристики текстов и редактировать их в диалоговом режиме с использованием словаря синонимов; производить лексический анализ текстов, при этом оценивать нагрузку на сенсорные каналы восприятия информации; настраиваться на лексически определенные группы людей посредством анализа характерных для них текстов.

Ценным является и отечественный программный продукт «Текстоанализатор», разработанный А. Ф. Толочко и Н. И. Миницким [5]. Среди функций программы можно отметить следующие: возможность точного математического описания авторского речевого стиля; наличие методов и процедур, позволяющих корректировать авторский стиль и обеспечить его единообразие по всему тексту; наличие технологии создания норм любых языков на основе кириллицы либо латиницы (белорусского, украинского, польского, английского и др.); оценка текста минимальных объемов и сравнение результатов этой оценки с образцами-эталонами; использование звуко-цветовых соответствий для проведения психолингвистической диагностики; проведение сравнительного анализа национальных учебников с аналогичными учебниками зарубежных стран.

Что касается компьютерных программ по изучению читабельности текста, следует отметить, что первые программы появились в начале 80-х гг. XX в.: Readability Calculations, Intext, Nisus Writer и др. Разработанные продукты предназначены для анализа английского, немецкого и других языков (но не русского).

Заключение. Исходя из вышеизложенного, можно сделать вывод, что в настоящее время отсутствуют исследования в области читабельности с использованием современных информационных технологий и необходимого инструментария для классификации русскоязычных текстов по ряду областей знаний в зависимости от подготовленности читателя. Это дает основание выделить ряд наиболее актуальных направлений, требующих детального изучения:

1. Исследование и разработка количественных критериев трудности понимания текста интересующей группой читателей. В этой связи проанализированы основные методы для определения трудности понимания различных текстов данной группой лиц и проведены эксперименты, которые позволили получить информацию относительно трудности текста в зависимости от подготовленности не только выбранной группы, но и потенциальных читателей [13].

2. Выбор структурных элементов исследуемых текстов, которые поддаются точному из-

мерению, и их детальное изучение. С этой целью следует использовать методы многомерного статистического анализа (кластерный и факторный анализ, метод корреляционных плеяд, многомерное шкалирование), которые позволят выявить связь между изучаемыми текстовыми признаками и на этой основе существенно сократить их количество.

3. Проведение дискриминантного анализа, который на основании наиболее информативных признаков текста позволит предсказать принадлежность объектов к двум непересекающимся группам, т. е. классифицировать исследуемые тексты в зависимости от их трудности для читателей. Результатом проведения дискриминантного анализа станет вывод дискриминантных функций, которые станут основой для разработки соответствующего программного инструментария для автоматической классификации текстов.

4. Создание компьютерной программы для классификации текстов в зависимости от трудности их восприятия читателями. Эта программа должна включать:

– поиск необходимых параметров текста и их вычисление;

– функции предварительной обработки, сохранения и загрузки данных;

– расчет на основе текстовых характеристик дискриминантных функций, необходимых для классификации текстов;

– принятие решения относительно трудности текста для потенциальных читателей.

Исследования по данным направлениям позволят поставить и в определенной степени решить вопрос о внедрении в редакционно-издательскую подготовку изданий автоматизированных систем, выполняющих информационные, логические, аналитические и другие задачи, решение которых до сих пор связывают иногда с деятельностью живого мозга. Полная или частичная замена человека (редактора) сложной специализированной системой позволит добиться не только невозможного для человека быстродействия, но и необходимого качества изданий благодаря объективной оценке трудности текста на основе его информационных характеристик.

Литература

1. Mendenhall, T. A. The characteristic curves of composition / T. A. Mendenhall // Science, 1887. – № 11. – P. 237–249.

2. Морозов, Н. А. Лингвистические спектры: средство для отличения плагиатов от истинных произведений того или иного неизвестного автора. Стилеметрический этюд / Н. А. Морозов // Известия отд. рус. языка и словесности Имп. Акад. наук. – Т. XX. – Кн. 4. – 1915.

3. Шевелев, О. Г. Разработка и исследование алгоритмов сравнения стилей текстовых произведений: автореф. дис. … канд. технич. наук / О. Г. Шевелев. – Томск, 2006. – 20 с.

4. Morton, A. Q. The authorship of greek prose / A. Q. Morton // Journal of the Royal Statistical Society (A), 1965. – № 128. – P. 169–233.

5. Миницкий, Н. И. Психолингвистические и информационные аспекты восприятия и обработки учебного текста / Н. И. Миницкий, А. Ф. Толочко // Белорус. психолог. журнал. – 2004. – № 3. – С. 57–61.

6. Закревская, Н. С. Исследование однородности текста с помощью модели скользящего среднего / Н. С. Закревская // Квантитативная лингвистика: исследования и модели: материалы Всерос. науч. конф., Новосибирск, 6–10 июня 2005 г. / НГПУ. – Новосибирск, 2005. – С. 26–33.

7. Хмелев, Д. В. Распознавание автора текста с использованием цепей А. А. Маркова / Д. В. Хмелев // Вестник МГУ. Сер. 9, Филология. – 2000. – № 2. – С. 115–126.

8. Хрулев, О. Определение автора по тексту на естественном языке [Электронный ресурс] / О. Хрулев. – Режим доступа: http://www.socionic.ru/articles/psycholingvist_author.htm, свободный.

9. Невдах, М. М. Формулы читабельности как критерий эффективного взаимодействия автора и читателя / М. М. Невдах // Научный потенциал студенчества – будущему России: материалы II Междунар. студ. конф., Ставрополь, 18–19 апр. 2008 г / СевКавГТУ. – Ставрополь: СевКавГТУ, 2008. – Т. 2: Лингвистика и межкультурная коммуникация. – С. 102–103.

10. Компьютерная обработка текстов при помощи ИС «СМАЛТ» / А. В. Король [и др.] // Проблемы развития гуманитарной науки на Северо-Западе России: опыт, традиции, инновации: материалы науч. конф., Петрозаводск, 29 июня – 2 июля 2004 г. / ПетрГУ. – Петрозаводск, 2004. – С. 122–124.

11. Гринбаум, О. Н. Проект «ЛИНДА» – автоматизированная система обработки лингвостатистических данных / О. Н. Гринбаум, Г. Я. Мартыненко, С. Я. Фитиалов // Прикладная лингвистика и автоматический анализ текста. – Тарту: ТГУ, 1988. – С. 31–33.

12. Система Polyanalist. Описание [Электронный ресурс]. – Режим доступа: http://www.megaputer.ru, свободный.

13. Невдах, М. М. Разработка количественных методов оценки трудности восприятия учебного текста для высшей школы / М. М. Невдах // Труды БГТУ. Сер. IX, Издат. дело и полиграфия. – 2008. – Вып. XVI. – С. 87–90.

Поступила 30.12.2008.