УДК 004.912

Г.Э. Романюк, М.С. Ходар

АНАЛИЗ ЛЕКСИКОГРАФИЧЕСКОЙ ОКРАШЕННОСТИ ХУДОЖЕСТВЕННЫХ ТЕКСТОВ

Предлагается программа анализа текста для нахождения и распознавания цветовых образов, используемых в тексте. С помощью программы фиксируются данные, строится матрица цветов и цветовые диаграммы. Полученные данные экспортируются в Excel для последующего изучения и анализа статистики цветовых образов и определения их зависимости от вида текста, жанра и автора.

Введение

Распознавание образов (объектов, сигналов, ситуаций, явлений или процессов) – едва ли не самая распространенная задача, которую человеку приходится решать практически ежесекундно от первого до последнего дня своего существования.

Сотни исследований и тысячи экспериментов в области изучения филологии, семантики и психологии были проведены за последние годы [1]. Было доказано множество разнообразных гипотез [2, 3]. В настоящей работе исследуется содержание различных цветовых образов в тексте и их влияние на сознание человека.

Теоретической базой работы является обширный список печатных изданий и интернет-ресурсов, посвященных описанию компьютерных программ, проблем компьютерной лингвистики, результатов исследований в области структурного литературоведения, существующим теориям цвета [4].

Объем информации, доступной в сети Интернет, растет с каждым годом. При этом большая часть этой информации представляет собой тексты на естественном языке. Научной новизной результатов настоящей работы является возможность извлечения цветовой информации из таких текстов и ее применение в научно-исследовательской и практической области при изучении влияния цвета на человека.

1. Основные методы анализа текстов

В разделе дается общее описание процесса анализа текста и рассматриваются основные проблемы [5]. На вход любой системы для анализа текста, как правило, поступает текст в исходном формате, т. е. в популярных форматах документов DOC, OAT, RTF, PDF, HTML. Такие документы кроме самого текста содержат также форматирование, сноски, «лишние» участки текста (рекламу). Первой сложностью при анализе текста является его извлечение из документов, поданных на вход. Далее текст следует пропустить через графематический анализатор, задача которого – определить границы слов, предложений, абзацев, а также отметить числа, знаки препинания, прочие символы и последовательности символов, не являющиеся словами, но играющие важную роль в предложении.

Когда предложения размечены и слова локализованы, начинает работу морфологический анализатор. Применяя имеющийся словарь и набор словоизменительных правил, этот анализатор дополняет каждое слово набором грамматических характеристик. На этапе морфологического анализа также возникает ряд сложностей с обработкой опечаток, особенно если они допущены в окончании слова, играющем ключевое значение в определении набора грамматических характеристик. Ряд сложностей возникает при анализе неизвестных слов, а также заимствованных из других языков.

Следующий этап анализа – синтаксический. Наиболее популярным подходом к этому анализу является построение дерева синтаксического разбора предложения. В процессе построения дерева исходное предложение перестраивается в древовидную структуру, представляющую собой аналог дерева синтаксического разбора некоторой формальной грамматики. При частичном

синтаксическом анализе полное дерево разбора не строится, а вместо этого анализатор сосредоточивается на поиске заранее определенных синтаксических конструкций в тексте.

Такой подход вполне приемлем, так как естественный язык не имеет грамматики в привычном математикам смысле. В естественных языках, как правило, достаточно вольные правила построения грамматических конструкций, слабо поддающиеся формализации. В достаточной степени формализуемы лишь некоторые синтаксические отношения:

– согласования и зависимости при построении словосочетаний;

– валентности глаголов и отглагольных причастий;

– ссылки на контекст (различные типы местоимений);

– общие правила построения предложений (простых предложений, сложносочиненных, сложноподчиненных и других сложных предложений).

Заключительным этапом анализа текста является семантический анализ результатов, полученных на этапе синтаксического анализа. Для этого этапа не существует устоявшихся моделей и подходов. В большинстве систем роль семантического анализатора играет эвристически реализованный модуль, осуществляющий поставленную перед ним задачу. Ясно, что даже при наличии дерева полного синтаксического разбора текста все равно сложно реализовать алгоритм, извлекающий полезную информацию на основе полученного дерева [1].

Одним из результатов исследования является новый подход к анализу текста – потоки интерпретаций. Потоки интерпретаций представляют собой множество пар (участок текста) и некоторую информацию. При этом не ставится никаких ограничений на тип информации и ее формат. Таким образом, основное отличие предлагаемого подхода от традиционного состоит в том, что если традиционно принято делить процесс анализа на этапы, жестко разделяя порядок исполнения этапов анализа, а также результат каждого этапа, то в подходе, основанном на потоках интерпретаций, предлагается фиксировать лишь формат результатов анализа.

Основные методы и методики анализа текстов:

Интент-анализ – метод, позволяющий реконструировать интенции автора по его тексту, поскольку для выявления и квалификации интенций опора на отдельные слова и предложения малопродуктивна [1]. Экспертное выявление и идентификация речевых интенций предоставляют 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: выделения круга обсуждаемых тем и вопросов, определения связей между объектами, затем кодификаций 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 влияния средств массовой коммуникации на индивидуальное и групповое сознание.

Ethnograph – методика, предназначенная для качественного (содержательного) анализа данных интервью, фокус-групп, дневников и пр. [1].

Процесс работы с данными построен на основе того, что осуществляется 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) и 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) – метод, позволяющий осуществлять контекстуальный анализ (анализ слов в пространстве четырех социальных контекстов: практики, традиций, эмоций, анализа) и анализ основных мыслей и образов, раскрытых в тексте. Для каждого вида анализа разработаны и стандартизированы нормы. Помимо указанных главных функций метод позволяет выводить статистику, проводить анализ слов, частотный анализ слов и категорий и т. д. [2].

Контент-анализ – самый распространенный метод, имеющий множество вариаций в различных методиках, позволяющий проводить 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 анализа 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 [1].

PROTAN – автоматизированная система контент-анализа, которая предназначена для анализа любого текста (рассказов, клинических интервью, научных публикаций, названий или резюме научных журналов, поэзии, рекламных материалов 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 [4].

Фоносемантический анализ текста (*или слова*) – метод, основанный на оценке его звучания безотносительно содержания. Заключается в сопоставлении системы сочетаний фонем в конкретном тексте или слове с их стандартизированными оценками по ряду биполярных шкал. 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 и методах [5]:

Vaal – методика, в которой реализованы алгоритмы оценки фонетического воздействия на человека слов и текстов русского языка, причем в основе эмоционального воздействия фоники слова и текста на подсознание человека лежат психофизиологические механизмы. Дает возможность 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 – 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. 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: фиксации изучаемого материала; выделении его формальных характеристик; обозначении контекста как коммуникативной ситуации; выборе направления и стратегии анализа; теоретическом дифференцировании и структурировании этапов исследования; определении техники и средств анализа при использовании конкретной модели исследования; дефиниции единиц анализа; проверке системы категорий в теории и на эмпирическом материале; осуществлении основных этапов исследования (дескрипции, реконструкции, интерпретации); фиксации результатов исследования, их обобщении, истолковании и структурировании. Дискурс-анализ позволяет выделить не только существенные характеристики социальной коммуникации, но и второстепенные, содержательные и формальные показатели (например, тенденции в вариа-

тивности 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, написанного на любом языке. Она основана 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 – метод обобщения прошлого опыта при помощи соотнесения последовательности слов в предложении и последовательности реальных (как предполагается) событий. 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 художественных и исторических текстов (мифов, легенд, летописей и пр.) до газетных статей, 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.), которые представляют 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, синтаксический анализ, интент-анализ и ряд других функций.

PC-ACE – методика, предназначенная для кодирования событий и позволяющая организовывать сложную текстовую информацию, хранить ее, осуществлять поиск необходимых данных, характеризующихся сложной структурой. Методика применима для всего спектра социальных наук и позволяет осуществлять качественный (содержательный) анализ данных.

Экспертная оценка текста – группа методов, в которую входят различные экспертизы текста, классификацию которых, согласно А.А. Леонтьеву [5], можно представить в следующем виде:

а) автороведческая экспертиза, направленная на установление автора текста или выявление категориальных признаков вероятного автора: пол, возраст, национальность, место рождения, место долговременного проживания, уровень образования и пр. ;

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: перефразирования текста или законченного фрагмента текста; семантического шкалирования, например методика семантического интеграла (В.И. Батов, Ю.А. Сорокин); свободного ассоциативного эксперимента; предикативного анализа текста. Существуют также компьютеризированные варианты данных экспертиз.

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, состоящего 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 примера можно отметить реализацию рассматриваемого 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 исследования текста, управления текстом, сравнения и пр.

Textanz – методика, предназначенная для частотного анализа текста на уровне слов, фраз, словоформ. С помощью данной методики можно осуществлять анализ любого текста. Она также применима для синтаксического анализа.

Цветонаименования в художественном тексте служат изобразительным средством. Классическая литература – достоверный источник сведений для понимания особенности цветов как эмоционально-перцептивных эталонов. Цвет в литературе – средство выражения, экспрессии. Он не только служит для моделирования абстрактного зрительного образа, но и сам встречается в определенных экспрессивных контекстах и жизненных ситуациях, которые определенным образом окрашены эмоционально и составляют органическую часть идейного содержания художественного произведения. Можно реконструировать внутреннюю форму цветов как символов эмоциональных состояний, изучая их употребление в произведении. Средством для такой реконструкции служат контекстные ситуации, в которые писатель вводит цвет как изобразительное средство.

2. Постановка задачи

Распознавание представляет собой задачу преобразования входной информации, в качестве которой уместно рассматривать некоторые параметры, признаки распознаваемых образов, в выходную, представляющую собой заключение о том, к какому классу относится распознаваемый образ.

В данном исследовании ставятся следующие задачи:

– разработать алгоритм и концепцию анализа текста на наличие цветовых образов;

– создать программу, которая сможет анализировать текст на содержание цветовых образов;

– научить программу отображать цветовую матрицу текста;

– разработать функцию выведения численной и графической статистики цвета в тексте;

– разработать методы применения программы для нахождения зависимостей между смыслом текста, его цветовым содержанием и влиянием на человека.

Успешность решения поставленных задач основывалась на наличии программно-технического обеспечения, определенной теоретической базы, перспективных идей и практических разработок.

3. Концептуальная модель разрабатываемой программы Text Analist

Цель информационного моделирования – создание концептуальной схемы предметной области. Эта схема (или просто модель) в упрощенном виде отражает наиболее важные для пользователей информационные объекты предметной области и связи между ними (рис. 1).

В разрабатываемом приложении пользователь является ключевым объектом системы, который управляет и инициирует процессы программы. Он должен самостоятельно выбирать и загружать исследуемый текст в программу. Именно пользователь решает, что именно следует проанализировать в тексте, и активирует те или иные функции программы. Все функции связаны с пользователем ассоциативной связью. Между двумя классами (объектами) существуют разные типы отношений. Самым базовым типом отношений является ассоциация (association). Это означает, что два класса связаны между собой. Обычно такое отношение используется на ранних этапах дизайна, чтобы показать, что зависимость между классами существует. Только ассоциативных связей недостаточно, чтобы на диаграмме классов показать все связи между объектами.

Классовые объекты «анализируемый текст», «цветовая диаграмма», «цветовая палитра», «матрица цветов» и Excel-файл связаны между собой такими связями, как агрегация и композиция. Обе они моделируют отношение «является частью» (HAS-A Relationship) и обычно выражаются в том, что класс целого содержит поля (или свойства) своих составных частей. Грань между ними достаточно тонкая, но важная (особенно в контексте управления зависимостями). Чтобы легче запомнить визуальную нотацию, ромбик всегда находится со стороны целого, а простая линия – со стороны составной части; закрашенный ромб означает более сильную связь – композицию, незакрашенный ромб показывает более слабую связь – агрегацию.

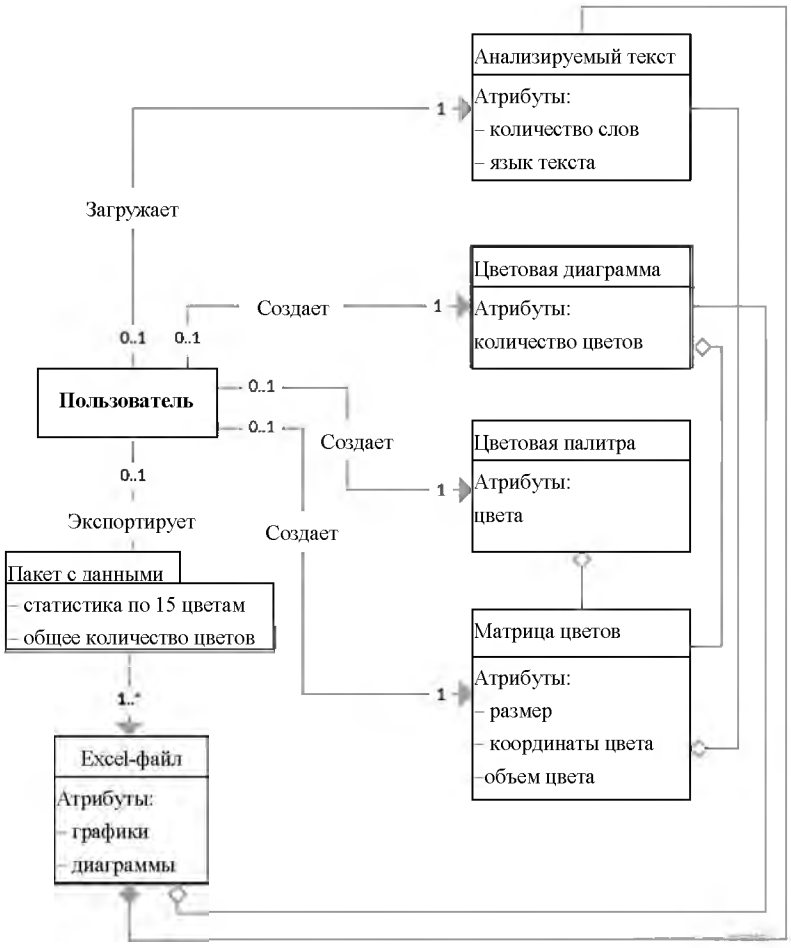

Рис. 1. Диаграмма классов

Разница между композицией и агрегацией заключается в том, что в случае композиции целое явно контролирует время жизни своей составной части (часть не существует без целого), а в случае агрегации не контролирует (например, составная часть передается через параметры конструктора):

```
        classCompositeCustomService
        {
//Композиция
        privatereadonlyCustomRepository_repository
=newCustomRepository();
publicvoidDoSomething()
{
//Используем_repository
        }
        }
```

```
        classAggregatedCustomService
        {
//Агрегация
        privatereadonlyAbstractRepository_repository;
publicAggregatedCustomService(AbstractRepositoryrepository)
{
_repository=repository;
}
publicvoidDoSomething()
{
//Используем_repository
        }
        }
```

CompositeCustomService для управления своими составными частями использует композицию, а AggregatedCustomService – агрегацию. При этом явный контроль времени жизни обычно приводит к более высокой связанности между целым и частью, поскольку используется конкретный тип, тесно связывающий участников между собой. С одной стороны, такая жесткая связь может не являться чем-то плохим, особенно когда зависимость стабильна. С другой стороны, можно использовать композицию и контролировать время жизни объекта без привязки к конкретным типам.

Концептуальная модель программы анализа текста в виде диаграммы классов помогает разработчикам на ранней стадии понять, какие объекты будут находиться в системе и какие классы следует создать при написании кода, что значительно сокращает время разработчика на понимание системы в целом. Существует несколько достаточно объективных критериев для определения связности дизайна по диаграмме классов: большие иерархии наследования (глубокие или широкие иерархии) и повсеместное использование композиции, а не агрегации говорят о сильно связанном дизайне.

Ключевой функцией разрабатываемого приложения является сравнительный анализ различных текстов с целью нахождения зависимостей и связей среди распознанных цветовых образов. Программа выполняет ряд основных и несколько второстепенных функций (рис. 2).

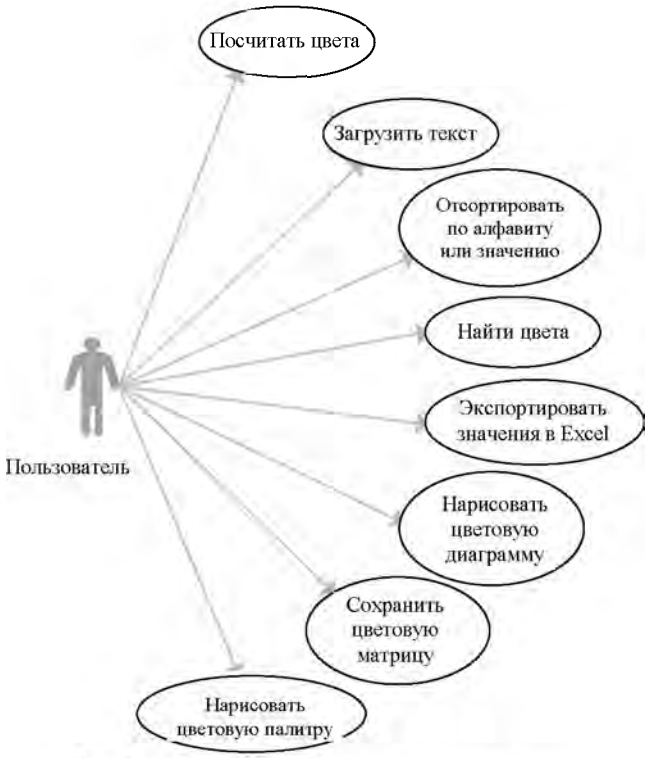

Рис. 2. Диаграмма вариантов использования

Основными функциями приложения Text Analist являются поиск цветов, подсчет их количества, вычисление объема, построение цветовой матрицы текста на основании этих данных и возможность статистического анализа полученных данных в программе Excel.

Второстепенными, вспомогательными функциями анализа являются построение цветовых диаграмм, цветовой палитры, поиск и подсчет цветов. Данные функции второстепенны, так как они не участвуют в решении задач, которые поставлены в настоящей работе, но они служат вспомогательным средством для визуализации, работы с данными и проверки полученных данных. Поэтому даже с учетом того, что эти дополнительные функции не являются обязательными, они значительно расширяют функциональность приложения для исследования текстов.

Функциональная спецификация требований к программному средству основана на диаграмме вариантов использования (см. рис. 2). Разрабатываемое программное средство распознавания цветовых образов на основе анализа текстов должно обеспечивать выполнение следующих функций:

– работу с пользовательским интерфейсом;
– анализ текстов;
– построение диаграмм;
– построение цветовой матрицы;
– импорт данных в Excel.

4. Описание программы Text Analist

Пользовательский интерфейс программы Text Analist (рис. 3 и 4) предназначен для анализа текстов на основе цветовых образов.

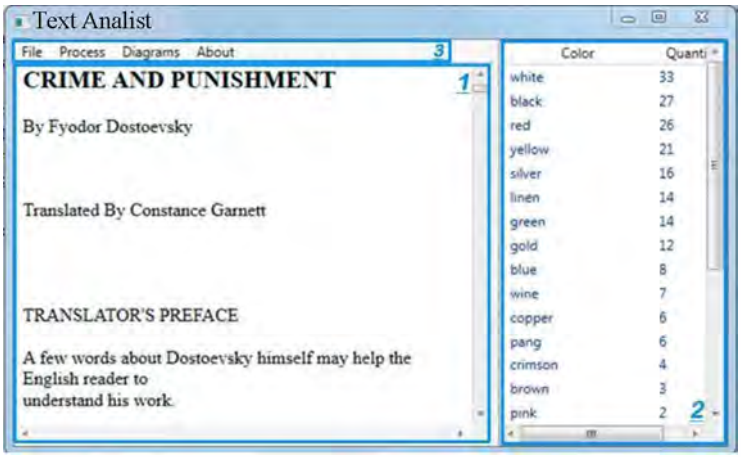

Рис. 3. Пользовательский интерфейс программы

Программа состоит из следующих основных блоков:

1 – текстовое поле, в котором содержится анализируемый текст;

2 – панель цветов, которая отображает, в каком количестве и какие цвета были найдены в тексте;

3 – панель инструментов, содержащая набор функций для анализа текста.

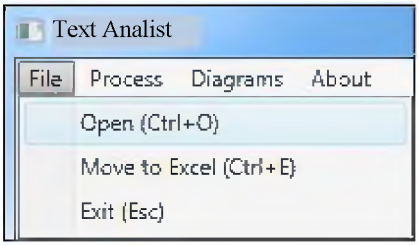

Рис. 4. Главное меню программы

Нажав на кнопку File -> Open, пользователь может выбрать нужный текстовый файл формата txt или RTF. Только после загрузки текста можно приступить к его анализу.

Для того чтобы выполнить поиск цветов в тексте, на панели инструментов необходимо нажать Process -> Find Colors. После этого появляется список цветов и их количество во втором блоке. Нажав на имя колонки, можно отсортировать данные по возрастанию или убыванию как по названиям цветов, так и по количеству их упоминаний (рис. 5).

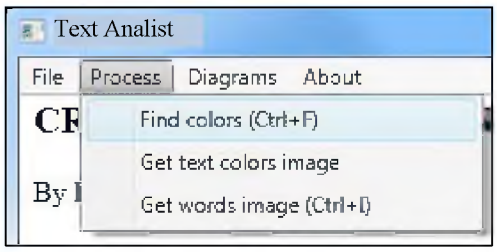

Рис. 5. Меню поиска текста и построения изображений

При нажатии Process -> Get words image программа анализирует текст и строит его цветовую матрицу. На примере (рис. 6) загружена книга Ф.М. Достоевского «Преступление и наказание». Видно, что чаще всего в книге упоминается белый цвет. Далее идут черный, красный, желтый и серебряный по частоте упоминания. Программа также не упускает такие цвета, как медный, цвет вина, малиновый закат, розовый, кремовый, горчичный, кукурузный, лимонный, лавандовый и т. д.

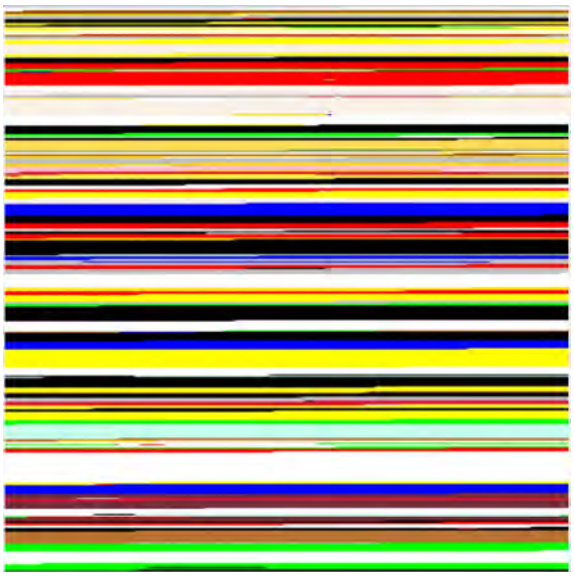

Рис. 6. Цветовая матрица романа Ф.М. Достоевского «Преступление и наказание»

На вход в программу подается текст в формате txt или RTF. Первым делом программа считает количество слов в тексте по количеству пробелов. Количество слов принимается за величину L. Извлекая из этого числа корень квадратный и округляя полученную величину до следующего целого числа, программа получает размер квадратной таблицы, которая заполняется цветами.

Приведем пример входного текста (рис. 7). Всего в тестовом тексте 100 слов. Это значит, что размер будущей таблицы будет $L = \sqrt{100} = 10$. Если количество слов в тексте, к примеру, 28 567, то размер таблицы будет $L = \sqrt{28567} \approx 169{,}01$. Округляем до следующего целого числа и получаем $L = 170$.

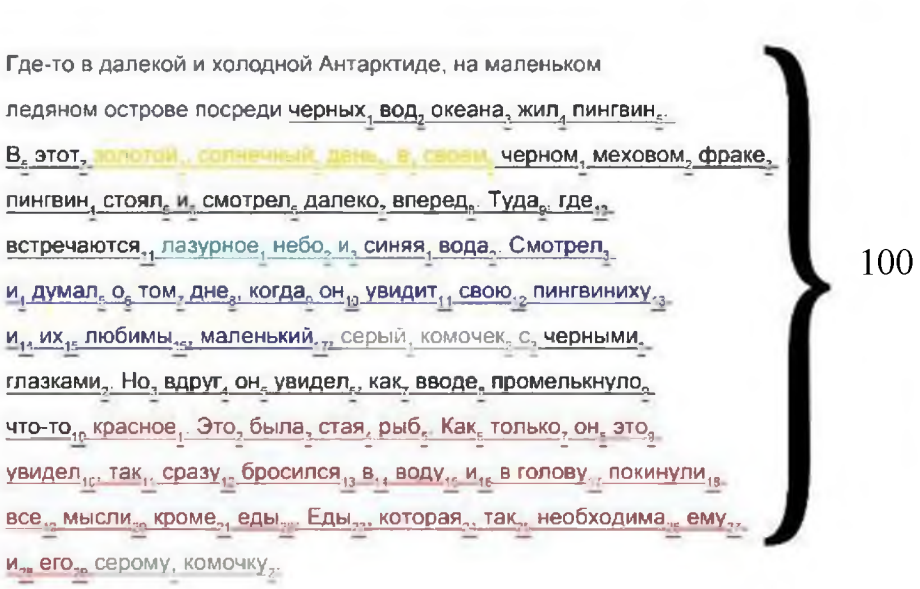

Рис. 7. Пример входного текста

На примере этого текста получается цветовая матрица в виде таблички (рис. 8).

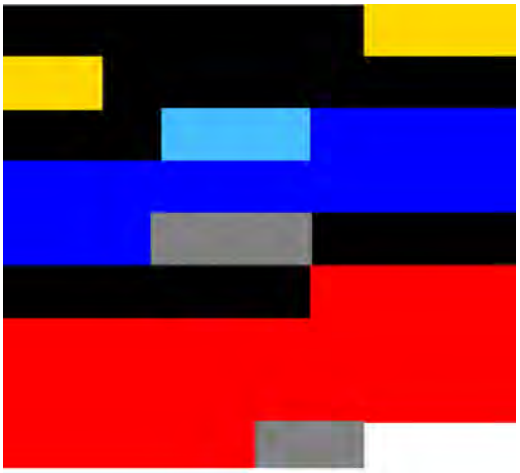

Рис. 8. Цветовая матрица входного текста

Алгоритм работы программы для создания цветовой матрицы следующий. Подразумевается, что текст уже проанализирован и слова-цвета в нем найдены. Первый цвет: «*черных* вод океана…» – рисуем первый черный квадратик. Далее, от слова *черных* до слова *золотой* семь слов. Значит, программа рисует семь черных квадратиков. Восьмой квадратик уже будет золотого цвета. От слова *золотой* до слова *черном* у нас еще пять слов. Значит, система рисует пять золотых клеточек. Далее идут 11 черных, 3 лазурных, 17 синих, 3 серых клеточки и так далее до конца текста. Аналогичным образом была получена палитра для текста романа «Преступление и наказание» (см. рис. 6).

Помимо цветовой палитры программа может строить круговые диаграммы. Нажимаем Diagrams -> Get find colors diagram и на выходе получаем диаграмму цветов 1 (рис. 9). На данной диаграмме все цвета разделены по секторам. При наведении на сектор курсора мышки выскакивает окошко с подсказкой, которое отображает количественную величину цвета. С помощью этой диаграммы можно наглядно увидеть, сколько и каких цветов было найдено в анализируемом тексте, и сравнить количество не только математически, но и визуально.

Нажимаем Diagrams -> Get images colors diagram и на выходе получаем диаграмму цветов 2 (рис. 10). Диаграмма 2 – это аналог цветовой матрицы, но в форме круговой диаграммы. Она показывает, в каком порядке и в каком объеме были найдены цвета в тексте.

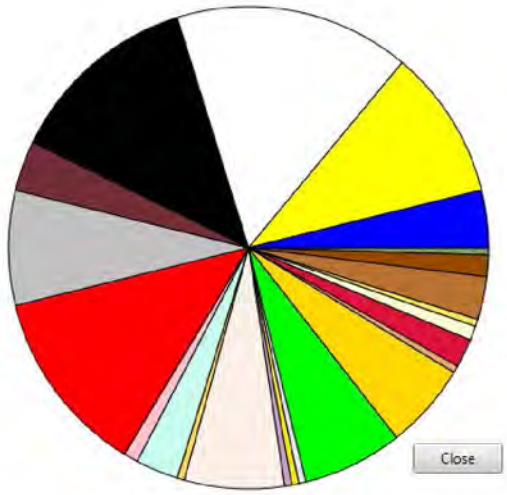

Рис. 9. Диаграмма цветов 1

Рис. 10. Диаграмма цветов 2

При нажатии File -> Move to Excel статистика по 15 самым популярным цветам переносится в файл Excel. Это специальный файл, в котором при построении таблицы графики строятся цветами, которые были выявлены ранее. Помимо этого переносится общее значение цветов в тексте. Таким образом, можно найти разницу между всеми цветами и 15 самыми популярными. Можно проводить сравнительный анализ данных сразу нескольких книг. Это могут быть книги одного жанра, одного автора и многих других направлений. В итоге данная программа не только анализирует текст и рисует его цветовую матрицу и палитру, но и дает возможность работать со статистическими данными с помощью средств Excel (рис. 11).

```
Sub SetChartColorsFromDataCells()

    If TypeName(Selection) <> "ChartArea" Then
        MsgBox "Сначала выделите диаграмму!"
        Exit Sub
    End If
    Set c = ActiveChart
    For j = 1 To c.SeriesCollection.Count
        f = c.SeriesCollection(j).Formula
        m = Split(f, ",")
        Set r = Range(m(2))

        For i = 1 To r.Cells.Count
            c.SeriesCollection(j).Points(i).Format.Fill.ForeColor.RGB = _
                r.Cells(i).Interior.Color
        Next i
    Next j
End Sub
```

Рис. 11. Макрос для закрашивания графиков и диаграмм

После закрытия Visual Basic можно вернуться в Excel (рис. 12).

Использовать созданный макрос очень просто. Запуск макроса происходит с помощью кнопки Макросы на вкладке Разработчик (Developer – Macros) или с помощью сочетания клавиш Alt+F8. В том же окне можно в случае частого использования назначить макросу сочетание клавиш с помощью кнопки Параметры (Options).

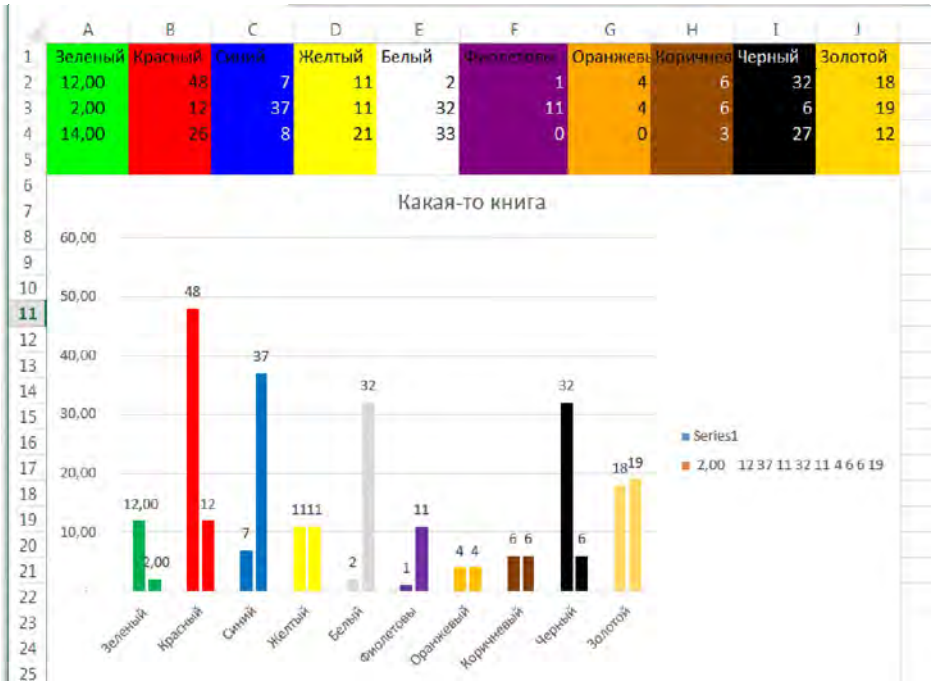

Рис. 12. Пример работы с созданным Excel-файлом

Разработанная система предназначена для распознавания образов на основе текстового содержания художественной литературы.

Заключение

Программа имеет две основные области применения: научно-исследовательскую и практическую. С точки зрения науки программа является средством для изучения и анализа влияния цвета на человека, его характер и психофизику путем разложения текстов на цветовые образы и сравнительного анализа числовых данных этих цветовых образов.

С практической точки зрения можно с помощью функции «цветовая матрица» создавать небольшие цветовые палитры, которые можно будет размещать на корешках книг как в обычных, так и интернет-магазинах. Благодаря «цветовой матрице» читатель сможет выбрать наиболее подходящую для него книгу на основании того, как выглядит рисунок, какие цвета и в каком объеме были использованы. Существует непосредственная зависимость между восприятием цветов книги и ее содержанием, сюжетом, историей, характером.

Помимо этого технологии данной программы можно применять в области психологии для изучения влияния книг и содержащихся в них цветов на подсознание человека и его внутренний мир.

Так как в настоящее время не существует непосредственных аналогов данного приложения, цветовой анализ может найти широкое применение в различных сферах человеческой жизнедеятельности.

Список литературы

1. Пескова, О.В. Алгоритмы классификации полнотекстовых документов / О.В. Пескова. – М. : МИЭМ, 2011. – 272 с.

2. Survey of Text Mining I: Clustering, Classification, and Retrieval / Ed. by M.W. Berry. – Springer, 2003. – 261 p.

3. Aggarwal, C.C. Mining Text Data / C.C. Aggarwal, C. Zhai. – Springer, 2012. – 527 p.

4. Пазельская, А. Метод определения эмоций в текстах на русском языке / А. Пазельская, А. Соловьев // The Intern. conf. on computational linguistics and intellectual technologies «Dialogue 2011». – М., 2011. – С. 510–522.

5. Жеребило, Т.В. Фоносемантика в словаре лингвистических терминов / Т.В. Жеребило. – 5-е изд. – Назрань : Пилигрим, 2010. – 486 с.

Поступила 21.09.2016

Белорусский национальный
технический университет,
Минск, пр. Независимости, 65
e-mail: galarom@tut.by

G.E. Romaniuk, M.S. Khodar

THE ANALYSIS OF LEXICOGRAPHICAL COLORING THE LITERARY TEXTS

A text analysis program for finding and recognition of color characters used in the text is developed. The program fixes the data, builds a color matrix and a color diagram. The program exports the data into Excel for further study and analysis of statistics of color characters to determine their dependence on text type, genre and author.