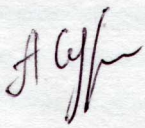
Нижегородский государственный технический университет

На правах рукописи



СУРКОВА Анна Сергеевна

РАЗРАБОТКА СТРУКТУРНО-СТАТИСТИЧЕСКИХ МЕТОДОВ И АЛГОРИТМОВ ИДЕНТИФИКАЦИИ ТЕКСТА

Специальность 05.13.01 «Системный анализ, управление и обработка информации»

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Нижний Новгород

2004

Работа выполнена на кафедре «Вычислительная техника» Нижегородского государственного технического университета

Научный руководитель: доктор технических наук,
профессор Л.С.Ломакина

Официальные оппоненты: доктор технических наук,
профессор А.Т.Надеев
кандидат физико-математических наук,
доцент А.Ф.Ляхов

Ведущая организация: Всероссийский институт научной
и технической информации
Российской академии наук (ВИНИТИ РАН)
г. Москва

Защита диссертации состоится «3» февраля 2005 года в 13[00] часов в аудитории 5427 на заседании диссертационного совета Д212.165.05 при Нижегородском государственном техническом университете по адресу: 603600, г.Нижний Новгород, ГСП-41, ул.Минина, 24, факс (8312)362311

С диссертацией можно ознакомиться в библиотеке Нижегородского государственного технического университета

Автореферат разослан «27» декабря 2004г.

Ученый секретарь
диссертационного совета К.т.н., доцент
Иванов А.П.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы.

Разработка и усовершенствование методов, направленных на автоматический анализ и автоматическую атрибуцию текстов разного уровня, приобретает все большую значимость на современном этапе и прикладной лингвистики, и текстологии, и автороведения в криминалистике, и других дисциплинах, связанных единым объектом исследования – текстом.

В связи с развитием электронных сетей и увеличением информации, распространяемой с их помощью, обостряется проблема соблюдения авторских прав. Традиционной для криминалистики была и остается задача идентификации автора анонимного текста по тем или иным его языковым параметрам (определение авторства различных анонимных писем, содержащих угрозы, шантаж и т.п.). Лингвистическая экспертиза важна при решении споров связанных с политическими проблемами. Анонимная или псевдоанонимная информация все чаще распространяется во время предвыборных кампаний с целью дискредитации конкурентов.

С другой стороны, начиная с первых попыток автоматизировать обработку текста, стало очевидно, что именно текст является наибольшей смысловой единицей языка. Именно текст как целостность, а не отдельные слова, предложения или абзацы необходимо изучать и обрабатывать с целью создания модели языка, адекватно отражающей значимые особенности естественных языков. Поэтому при проектировании автоматических систем обработки естественного языка необходимо учитывать те особенности текста, которые отражают его системные свойства. Изучение структуры целого текста является необходимой базой для дальнейших исследований в этом направлении и реализации результатов при создании систем автоматической обработки текстов. Однако ограниченность вычислительных ресурсов и недостаточная разработанность теоретической базы привели к тому, что известные в настоящее время алгоритмы автоматической обработки текстов носят, как правило, частный характер и разрабатываются для каждой конкретной задачи.

Степень разработанности проблемы.

Разработкой проблем, связанных с задачами автоматической обработки текстов, в последние годы активно занимаются ученые в нашей стране и за рубежом. Современные работы основываются на результатах, полученных в процессе становления автоматической обработки текстов, как особого направления компьютерной лингвистики, в работах Р.Г.Пиотровского, И.П.Севбо, А.А.Поликарпова, Ю.А.Шрейдера, М.В.Арапова, Б.В.Сухотина. Вопросы построения общих систем АОТ и систем идентификации текстов рассматривались в работах таких зарубежных авторов как Г. Йеля, Д. Ципфа, Г. Хьетсо, В. Фукса, Д. Холмса, Д. Барроуза, Ф. Твиди.

В настоящее время проблемами атрибуции и установления авторства занимаются такие авторы, как М.А.Марусенко, Г.Я.Мартыненко, О.В.Кукушкина, Л.И. Бородкин.

Среди работ в области юридической лингвистики можно отметить работы Н.Д.Голева, Е.И.Галяшиной, А.Ю.Комиссарова. Однако большинство подобных работ носят прикладной характер, но прикладные исследования не предоставляют систематизированной теоретической базы лингвистических знаний, позволяющей однозначно решать вопросы спорного авторства.

В последнее время стала очевидной необходимость системного рассмотрения совокупности текстов разных авторов, стилей и жанров; в связи с этим тема диссертационной работы является актуальной, как для теоретических, так и для практических исследований.

Цель работы.

Целью работы является построение модели текста как системы, установление структурных инвариантов текста различного уровня и на их основе разработка методов и алгоритмов идентификации текстов.

Задачи работы.

Для достижения намеченной цели требуется решение следующих основных задач:

- Построение структурно-иерархической модели текста.

- Разработка алгоритма статистической обработки текста с целью выявления различных параметров, характеризующих структуру текста.
- Проверка возможности использования некоторых структурных параметров в качестве инвариантов текста различного уровня.
- Разработка метода объединения результатов идентификации разными методами.

Объект исследования.

В качестве объекта исследования рассматривались тексты на русском языке различных авторов XIX-XX веков; тексты с различной жанрово-стилевой принадлежностью: художественные, научные, публицистические.

Методы исследований

Методологической основой данной работы является системный анализ. Для теоретических исследований применялись методы теории вероятностей, математической статистики, теории информации.

Научная новизна.

На основании исследования текста как системы разработан новый метод установления авторских инвариантов текста.

Предложена методика идентификации текстов на основе полученных авторских инвариантов.

Разработаны алгоритмы сравнения структур текстов на основе сравнения сечений многомерных законов распределения букв в словах.

Обоснованность и достоверность результатов работы.

Обоснованность и достоверность результатов обеспечены корректным использованием в работе современного математического аппарата и подтверждены результатами экспериментальных исследований конкретных текстов.

Практическая значимость.

Практическая ценность заключается в возможности применения полученных результатов в задачах поиска информации, при создании

информационно-поисковых систем, при проведении автороведческой экспертизы, при установлении спорного авторства.

Реализация результатов работы.

Разработанные в рамках диссертационной работы алгоритмы анализа структуры текста и методика идентификации текстов используются в учебном процессе Нижегородского государственного лингвистического университета им. Н.А. Добролюбова и Нижегородской Академии МВД России.

Апробация результатов работы.

Основные положения и результаты работы представлялись и докладывались на следующих научных конференциях:

- Международной конференции «Математика. Образование. Гендерные проблемы.» (Воронеж, 2000);
- 6-ой международной конференции «НТИ-2002. Информационное общество. Интеллектуальная обработка информации. Информационные технологии.» (Москва, ВИНИТИ, 2002)
- Всероссийской научно-технической конференции. «Информационные системы и технологии ИСТ-2003.» (Н.Новгород, НГТУ, 2003).
- 2-ой региональной научно-технической конференции «Будущее технической науки Нижегородского региона». (Н.Новгород, 2003).
- 3-ей научно-технической конференции «Будущее технической науки Нижегородского региона». (Н.Новгород, 2004).
- Всероссийской научно-технической конференции. «Информационные системы и технологии ИСТ-2004.» (Н.Новгород, НГТУ, 2004).
- Всероссийской научно-методической конференции «Языковые и культурные контакты различных народов» (Пенза, 2004).

Публикации.

По теме диссертационной работы опубликовано 10 работ.

Структура и объем диссертации.

Диссертация состоит из введения, четырех глав, заключения, списка литературы и приложений.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении дана общая характеристика работы, обоснована актуальность выбранной темы, сформулированы цели и задачи исследования, показана научная новизна и практическая ценность работы. Кратко изложено содержание диссертации по главам.

В первой главе приведен обзор существующих методов и принципов автоматической обработки текстов, применяемых для различных целей – информационный поиск, автоматическое индексирование и реферирование, идентификация текстов, судебная автороведческая экспертиза.

В разделе 1.1. рассматриваются некоторые действующие системы автоматического анализа текстовой информации, основанные на различных представлениях текста. Здесь же определены основные проблемы, возникающие при реализации тех или иных методов обработки текста. В качестве значимых причин указывается недостаточно разработанная теоретическая база общей лингвистики текста, сложность формализации естественных языков. При создании сложных систем обработки больших текстовых массивов необходима формализация представления текста на всех уровнях его представления: морфологическом, лексическом, синтаксическом, грамматическом, семантическом. В теоретической литературе последнего времени идет активный поиск возможностей представления знаний, разрабатываются различные принципы интерпретации смысла текста и формализации семантического уровня.

В разделе 1.2. рассматриваются особенности применения методов компьютерной лингвистики, разработанных в целях атрибуции текстов, для задач криминалистики. Особое междисциплинарное направление, которое занимается применением лингвистических познаний в юридической области, получило название юридической лингвистики. Лингвистические методы определения авторства используются при проведении судебных автороведческих экспертиз (при рассмотрении письменных текстов, выполненных анонимно или когда авторство спорно) а также фоноскопических экспертиз в качестве составной части комплексных методик идентификации.

В разделе 1.3. рассматриваются основные принципы и положения одной из актуальных задач автоматической обработки текстов – задачи атрибуции текста. Под атрибуцией понимается не только определение его автора, но соотнесение тексту соответствующих ему атрибутов, к которым причисляется имя создателя, жанр произведения, время и место создания текста. В табл. 1. представлены основные методы определения авторства текста.

Таблица 1. Обзор основных методов определения авторства текста.

№ п/п	Название метода	Авторы	Основные особенности метода	Недостатки
1	Атрибуция на основе лексического уровня языка	Г.Хьетсо, С.Густавссон, Б.Бекман	В качестве параметров, используются: длина предложения, длина слова, богатство словарного запаса и другие	Не учитываются структурно-синтаксические параметры, которые более полно определяют стиль.
2	Анализ «графов зависимостей»	И.П.Севбо, Ю.И.Петунин, Е.Д.Галюта	Метод основан на анализе графов синтаксических связей типичных предложений	Недостаточная разработка теоретических методов для автоматизации предварительной обработки текстов (составление деревьев)
3	Методы распознавания образов в целях атрибуции текстов	М.А.Марусенко, А.А.Рогов, Г.Я.Мартыненко	Текст описывается с помощью 112 параметров, из них выбираются наиболее информативные для рассматриваемых авторов, и методами распознавания образов производится идентификация.	Значительное количество признаков Необходимо определять «вручную». Нет возможности рассматривать большое количество авторов.
4	Анализ частот парных встречаемостей грамматических классов	Л.В.Милов, Л.М.Бородкин,	Сравнение текстов производится по наиболее вероятным для данных авторов парным встречаемостям грамматических классов	Трудоемкая предварительная обработка текстов пока слабо автоматизирована.
5	«Лингвоанализатор», «Атрибутор»	ПоликарповА.А. Хмелев Д.В. Тимашев А.Н.	В качестве признаков для анализа и оценки индивидуального авторского стиля используются сочетания двух букв в слове (пары) и трех букв (триады).	Не рассматриваются произведения разных стилей писателей, идентификация проводится по одному параметру.

Общим недостатком всех методов является стихийность выбора параметра, по которым производится идентификация. Все методы кроме последнего предполагают участие человека в предварительной обработке текста, поэтому нет возможности производить исследование на большом количестве авторов.

Во второй главе обосновывается необходимость рассмотрения текста, как системы, и предлагаются некоторые модели структурирования текста.

В разделе 2.1. рассматриваются основные особенности системного подхода к анализу текстов.

В разделе 2.2 на основании системного подхода построена структурно-иерархическая модель текста, которая представлена на рис.1.

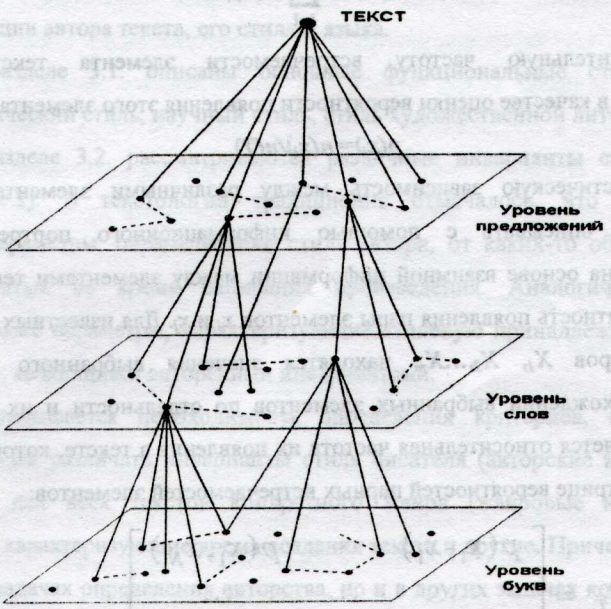


Рис.1Структурно-иерархическая модель текста.

В иерархии можно выделить уровни букв, слогов, слов, предложений. Связи между элементами нижнего уровня регламентируются элементами высшего уровня.

В разделе 2.3. приведены основные модели текста, основанные на учете различных структурно-статистических характеристик текста. К таким моделям

относится структурно-вероятностная модель, представление текста с использованием цепей Маркова, информационная модель.

Рассмотрим текст в виде последовательности дискретных случайных событий (последовательность появления лингвистических единиц: букв, слогов, словоформ и т.д.). Пусть x_i, $i=1,2...N$ - элемент текста, N- число различных значений, которые может принимать элемент x_i. Для каждого элемента x_i можно указать целое число $n(x_i)$, которое будет характеризовать частоту употребления элемента в тексте, назовем эту величину встречаемостью элемента x_i в тексте. Общая сумма встречаемости всех элементов равна объему этого текста $n(T)$.

$$n(T) = \sum_{i=1}^{N} n(x_i)$$

Относительную частоту встречаемости элемента текста можно использовать в качестве оценки вероятности появления этого элемента в тексте:

$$p(x_i) = n(x_i)/n(T)$$

Статистическую зависимость между различными элементами текста предлагается описывать с помощью информационного портрета текста, строящегося на основе взаимной информации между элементами текста. Пусть $p(x_i,x_j)$ - вероятность появления пары элементов x_i и x_j. Для известных текстов T_1, $T_2,...T_n$ авторов X_1, $X_2,...X_m$ находятся значения выбранного параметра: количество вхождений выбранных элементов по отдельности и их сочетаний, затем вычисляется относительная частота их появления в тексте, которую можно записать в матрице вероятностей парных встречаемостей элементов:

$$B_{T_k} = \begin{bmatrix} p(x_1,x_1) & \cdots & p(x_1,x_N) \\ \cdots & \cdots & \cdots \\ p(x_N,x_1) & \cdots & p(x_N,x_N) \end{bmatrix}, \; k=1,2,...n$$

Тогда каждой паре элементов может быть поставлена в соответствие количественная мера взаимной информации между ними и результаты могут быть представлены в виде матрицы взаимной информации между элементами названа информационного портрета текста:

$$A_{T_k} = \begin{bmatrix} a_{11} & \cdots & a_{1N} \\ \cdots & \cdots & \cdots \\ a_{N1} & \cdots & a_{NN} \end{bmatrix}, k=1,2,\ldots n,$$

где $a_{ij}=I(x_i,x_j)$ - взаимная информация между элементами x_i и x_j, вычисляемая по формуле:

$$I(x_i, x_j) = \log \frac{p(x_i, x_j)}{p(x_i)\,p(x_j)}, i,j=1,2,\ldots N. \qquad (1)$$

В третьей главе представлены разработанные методы и алгоритмы идентификации автора текста, его стиля и языка.

В разделе 3.1. описаны основные функциональные стили текстов (публицистический стиль, научный стиль, стиль художественной литературы).

В разделе 3.2. рассматриваются различные инварианты стиля, автора текста, (рис.2). В текстологии традиционно отмечалось, что необходимо различать параметры, определяющие стиль автора, от каких-то общеязыковых норм, принятых во время написания произведения. Аналогично, следует различать также параметры, характеризующие стилевую принадлежность текста и параметры, являющиеся авторскими инвариантами.

Обосновывается необходимость определения критериев, по которым можно было бы различать инварианты стиля писателя (авторские инварианты), инварианты для всех текстов конкретного жанра (жанровые инварианты), инварианты, характеризующие время создания текста и другие. Причем это важно не только в задачах определения авторства, но и в других задачах компьютерной лингвистики и автоматической обработки текста. Например, при создании систем информационного поиска возникает необходимость автоматического разделения текстов по стилям или времени создания, что требует определения четких формализованных критериев стилей текстов.

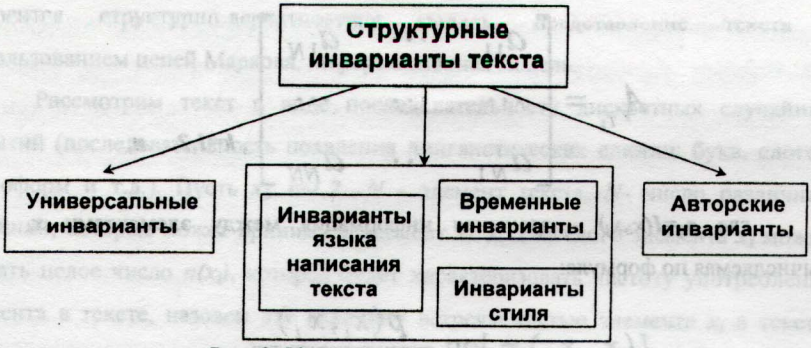


Рис. 2. Уровни инвариантов текста.

В разделе 3.3. описываются методы построения информационных портретов текстов, основанных на некоторых характеристиках текста, отражающих его внутреннюю структуру. Структура языка писателя начинает проявляться в связях между элементами текста низшего уровня, то есть связь между словами в предложении отражена в структуре слова. В рамках теории вероятностей структура слова описывается многомерным законом. Но при сравнении многомерных законов возникают проблемы на этапе реализации. Поэтому предлагается сравнивать отдельные сечения многомерного закона. К таким сечениям можно отнести двумерные законы распределения различных сочетаний букв в словах.

Для сравнения текстов предлагается использовать коэффициент корреляции K и среднеквадратическое отклонение σ^2, вычисленные по информационным портретам сравниваемых текстов. Информационные портреты строятся по некоторым сечения многомерного закона распределения букв в словах (отдельные буквы на заданном расстоянии в тексте, пары и триады букв и т.д.). При сравнении отдельных текстов рассматриваются информационные портреты этих текстов, а при сравнении авторского стиля писателя или функциональных стилей предварительно вычисляется обобщенная информационная матрица. Если $a_{ij}^{(1)}$, $a_{ij}^{(2)}$ - элементы двух сравниваемых матриц информации, то коэффициент корреляции вычисляется по формуле:

$$K = \frac{\sum_{i,j} a_{ij}^{(1)} a_{ij}^{(2)}}{\sqrt{\sum_{i,j} (a_{ij}^{1})^2} \sqrt{\sum_{i,j} (a_{ij}^{2})^2}} \qquad (2)$$

А среднеквадратическое отклонение:

$$\sigma^2 = \frac{\sum_{i,j} \left(a_{ij}^{(1)} - a_{ij}^{(2)} \right)^2}{q}, \qquad (3)$$

q – число пар в сравниваемых матрицах.

Близость текстов друг к другу определяется наибольшим значением коэффициента корреляции и наименьшим значением σ^2.

В разделе 3.4. рассматриваются некоторые структурные характеристики текста, и исследуется возможность применения их в качестве инвариантов разного уровня.

Были рассмотрены следующие характеристики текста:

- пары букв, идущие подряд в слове
- две буквы, встречающиеся в слове через одну букву
- пары двухбуквенных сочетаний, идущие в слове подряд. (из всех возможных 1024=32*32 двухбуквенных сочетаний для исследования были выбраны 30 наиболее часто встречаемых в русских текстах: ва, ка, ла, на, ра, та, ов, не, ре, ли, ни, ал, ел, ол, ен, он, во, го, ко, ло, но, по, ро, то, ер, ор, пр, ат, от, ст)
- пары гласных букв (в слове не обязательно стоящие рядом)
- пары служебных слов в тексте

Идентификация по двум последним параметрам (гласным буквам в слове и парам служебных слов в тексте) производится плохо (около 15% контрольных текстов идентифицировано верно), поэтому эти параметры исключили из дальнейшего рассмотрения.

В разделе 3.5. описывается обобщенный алгоритм идентификации. Если было рассмотрено ***s*** двумерных законов распределения, то коэффициент корреляции и среднеквадратическое отклонение можно интерпретировать как координаты в Евклидовом пространстве размерности ***2s***. Каждому тексту будет соответствовать точка в этом пространстве параметров. Текстам, принадлежащим одному автору, соответствует компактное множество точек в пространстве.

Для каждой точки можно не указывать конкретное значение каждой координаты, а важно знать расстояние между объектами. Каждое значение коэффициента корреляции и среднеквадратического отклонения характеризует расстояние между текстами по какой-либо координате. Для вычисления общего расстояния используется формула расстояния в евклидовом пространстве:

$$L = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2},$$

где ***n*** - размерность пространства, x_i, y_i, ***(i=1,2,…n)*** - координаты точек. Поскольку величины коэффициента корреляции и среднеквадратического отклонения характеризуют расстояние между сравниваемыми объектами, то расстояние вычисляется по формуле:

$$L = \sqrt{\sum_{i=1}^{s} (1 - K_i)^2 + \sum_{i=1}^{s} \sigma_i^2} \qquad (4)$$

Решение о принадлежности исследуемого текста какому-либо автору из обрабатываемого списка принимается по минимальному расстоянию L.

В четвертой главе приведены примеры практического применения предложенных методов идентификации текста.

В разделе 4.1. описаны результаты идентификации, проведенной по совокупности характеристик. В работе были обработаны произведения 20 авторов XIX-XX веков. Для контроля предъявлялись контрольные тексты (по 1-2) каждого автора. Результаты идентификации представляются в виде таблиц. В табл. 2 приведен фрагмент такой таблицы для авторов первой половины XX века.

Информационные портреты были построены на основе вероятностей встречаемостей соседних букв в слове. В таблице для каждого контрольного произведения вычисляются значения *(1-K)* и σ^2 и выделяются цветом минимальные значения в строке, т.е. определяется наиболее вероятный автор. Как видно из таблицы, три произведения «Коновалов» и «Челкаш» Горького и «Другие берега» Набокова идентифицированы неверно.

Таблица 2. Значения коэффициента корреляции и σ^2

Корреляция и Среднеквадратическое отклонение (одиночные буквы)							
произведение:		Булгаков	Горький	Грин	Набоков	ТолстойАН	Фадеев
Собачье сердце	1-K	0,078266	0,103656	0,106095	0,091668	0,096947	0,097918
---	σ^2	0,291424	0,486525	0,41967	0,3488	0,366212	0,379455
Коновалов	1-K	0,089882	0,076056	0,074777	0,076578	0,08043	0,081951
---	σ^2	0,274755	0,261897	0,231955	0,230799	0,242819	0,261296
Челкаш	1-K	0,115134	0,101117	0,111817	0,103529	0,094163	0,118473
---	σ^2	0,276479	0,264766	0,27316	0,248493	0,226234	0,288793
Золотая цепь	1-K	0,075911	0,107005	0,042992	0,073684	0,076921	0,087206
---	σ^2	0,285409	0,476625	0,167952	0,277839	0,290772	0,339839
Другие берега	1-K	0,131074	0,17945	0,135487	0,108786	0,128201	0,135071
---	σ^2	0,266185	0,494582	0,325527	0,227412	0,270154	0,314243
Бледное пламя	1-K	0,062201	0,098659	0,067502	0,05116	0,07418	0,073312
---	σ^2	0,336089	0,69048	0,380547	0,271271	0,40708	0,438527
Буратино	1-K	0,119269	0,14711	0,121783	0,103303	0,09848	0,136316
---	σ^2	0,344815	0,4931	0,377435	0,304153	0,281941	0,403194
Хождение (книга 2)	1-K	0,04932	0,068747	0,074545	0,063539	0,039069	0,05866
---	σ^2	0,250659	0,448871	0,381335	0,312755	0,203946	0,315492
Фадеев_ Разгром	1-K	0,080301	0,063465	0,077708	0,084747	0,073783	0,058373
---	σ^2	0,343062	0,296348	0,332444	0,354931	0,320994	0,260666

Аналогично строятся таблицы значений коэффициента корреляции и среднеквадратического отклонения, полученные при сравнении информационных портретов, вычисленных по другим характеристикам (встречаемости букв через одну в слов и встречаемостей в слове наиболее вероятных двухбуквенных сочетаний).

Рассматривая *(1-K)* и σ^2 в качестве расстояний между объектами, по формуле (4) вычисляется «расстояние» между контрольными произведениями и каждым автором. Значения записываются в таблицу, фрагмент которой представлен в табл. 3. Цветом выделены клетки, соответствующие трем наиболее вероятным автора для каждого произведения (интенсивность цвета уменьшается с вероятностью).

Таблица 3.

произведение:	Булгаков	Горький	Грин	Набоков	ТолстойАН	Фадеев
Собачье сердце	0,63502	0,945618	0,804861	0,72212	0,71846	0,7875
Коновалов	0,668948	0,596687	0,668137	0,596809	0,659826	0,643247
Челкаш	0,616218	0,663268	0,62668	0,610053	0,592153	0,604747
Золотая цепь	0,701402	0,855219	0,40034	0,574467	0,687239	0,79562
Другие берега	0,8409	1,152317	0,891057	0,690288	0,770833	0,832931
Бледное пламя	0,707579	1,123716	0,712661	0,515346	0,755451	0,863748
Буратино	0,765641	0,973014	0,901919	0,769712	0,663935	0,875781
Хождение (книга 2)	0,643789	0,9158	0,803442	0,721134	0,459283	0,638962
Фадеев_ Разгром	0,721919	0,674544	0,668037	0,707864	0,671025	0,484278

При объединении результатов различных идентификаций определение автора стало точнее. Произведения «Коновалов» и «Другие берега» отнесены к их настоящим авторам, и только «Челкаш» был идентифицирован неверно.

В целом были проверены 32 произведения 20 авторов. В результате обобщенной идентификации верно было идентифицировано 28 произведений (88%), в двух произведениях (6%) истинный автор был поставлен на второе или третье место. Результаты приведены на диаграмме 1.

Диаграмма 1.

В разделе 4.2. описано применение разработанного алгоритма для определения стиля произведения и языка.

В разделе 4.3. рассматривается возможность применения предложенных методов для определения автора предложенного текста при проведении автороведческой экспертизы.

В заключении изложены основные результаты диссертационной работы.

В приложении приведены программные продукты, реализующие предложенные алгоритмы, результаты статистической обработки текстов и документы, подтверждающие практическое применение результатов диссертационной работы.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

1. В результате анализа научной литературы установлено, что существует проблема более полного использования возможностей статистической обработки текста с целью его идентификации.
2. Рассмотрение текста через структурно-иерархическую модель, позволило выявить новые свойства текста, которые являются устойчивым проявлением индивидуальных особенностей автора в разной степени на всех иерархических уровнях.
3. Установлено, что индивидуальные особенности автора проявляются как авторские инварианты текста на уровне букв и их сочетаний в виде соответствующих многомерных законов распределения.
4. Разработаны методы и алгоритмы идентификации текстов посредством сравнения двумерных законов распределения букв и их сочетаний (информационных портретов), полученных из многомерных законов.
5. Разработаны методы и алгоритмы идентификации текстов на основе объединения результатов сравнения отдельных информационных портретов.
6. В результате практических исследований подтверждена эффективность предложенных методов и алгоритмов (из контрольных текстов верно идентифицировано около 90%)
7. Разработанные методы идентификации текстов могут служить основой для дальнейших модификаций и вариантов алгоритмического и программного обеспечения автоматизированных систем поиска информации.

ПУБЛИКАЦИИ

Основное содержание диссертации отражено в следующих публикациях:

1. Суркова А.С. Определение инвариантов разного уровня в задачах атрибуции текстов. //Языковые и культурные контакты различных народов: Сборник статей Всероссийской научно-методической конференции. - Пенза, 2004. с. 251-252.

2. Суркова А.С. Построение системных инвариантов текста и его идентификация. //Материалы Всероссийской научно-технической конференции. ИСТ-2004. – Н.Новгород, НГТУ, 2004. с.122

3. Суркова А.С. Проблема идентификации автора текста в юрислингвистике. //Материалы 3-ей молодежной научно-практической конференции «Будущее технической науки» – Н.Новгород, 2004. с.51-52.

4. Панкратова А.З. Суркова А.С. От текста к информационному портрету. //Материалы 2-ой региональной научно-технической конференции «Будущее технической науки Нижегородского региона». – Н.Новгород, 2003. с.37.

5. Ломакина Л.С., Панкратова А.З. Суркова А.С. Автоматический анализ большого текстового массива. //Материалы 2-ой региональной научно-технической конференции «Будущее технической науки Нижегородского региона». – Н.Новгород, 2003. с.33.

6. Ломакина Л.С., Панкратова А.З. Суркова А.С. Развитие методов анализа и оптимизации структуры текста с целью идентификации. //Материалы Всероссийской научно-технической конференции. ИСТ-2003. – Н.Новгород, НГТУ, 2003. с.138.

7. Ломакина Л.С., Суркова А.С. Системный подход в лингвистических исследованиях. //Материалы 6-ой международной конференции «НТИ-2002. Информационное общество. Интеллектуальная обработка информации. Информационные технологии.» - М.: Изд-во ВИНИТИ, 2002. с.224-225.

8. Голубева(Суркова) А.С., Ломакина Л.С. Разработка алгоритма кластерного анализа в лексико-семантических исследованиях. //Межвуз. сб.

«Системы обработки информации и управления». Выпуск 7. – Н.Новгород, 2001. с.41-44.

9. Голубева(Суркова) А.С., Ломакина Л.С. Применение методов многомерного статистического анализа в лексико-семантических исследованиях. //Материалы Всероссийской научно-технической конференции, посвященной 65-летию информационных систем и технологий НГТУ. ИСТ-2001. – Н.Новгород, НГТУ, 2001. с.175-176.

10.Голубева(Суркова) А.С. Применение методов кластерного анализа в исследованиях лексико-семантических групп. //Материалы международной конференции «Математика. Образование. Гендерные проблемы.». Том 1. – Воронеж: Изд-во НОУ «Интерлингва», 2000. с.82-83.

Подписано в печать 23.12.04. Формат 60 × 84 1/16. Бумага офсетная.
Печать офсетная. Уч.-изд. л. 1,0. Тираж 100 экз. Заказ 837.

Нижегородский государственный технический университет.
Типография НГТУ. 603600, Нижний Новгород, ул. Минина, 24.