На правах рукописи

Панкратова Анна Зурабовна

РАЗРАБОТКА МОДЕЛИ И МЕТОДА СТРУКТУРИРОВАНИЯ ТЕКСТА С ЦЕЛЬЮ ЕГО ИДЕНТИФИКАЦИИ

Специальность 05.13.17

Теоретические основы информатики

(технические науки)

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Нижний Новгород

2002

На правах рукописи

Панкратова Анна Зурабовна

РАЗРАБОТКА МОДЕЛИ И МЕТОДА СТРУКТУРИРОВАНИЯ ТЕКСТА С ЦЕЛЬЮ ЕГО ИДЕНТИФИКАЦИИ

Специальность 05.13.17

Теоретические основы информатики

(технические науки)

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Нижний Новгород

2002

743374

Работа выполнена на кафедре вычислительной техники Нижегородского государственного технического университета

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

Официальные оппоненты. доктор технических наук, профессор В.А. Утробин

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 «20» июня 2002 г. в 14 часов на заседании диссертационного совета Д212.165 05 при Нижегородском государственном техническом университете

С диссертацией можно ознакомиться в библиотеке Нижегородского государственного технического университета

Автореферат разослан «13» мая 2002 г.

Ученый секретарь диссертационного совета А.П. Иванов

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы.

Новые информационные технологии связаны с созданием и обработкой большого количества разнородных текстов. Необходимость обработки все увеличивающегося количества текстов требует разработки новых и модификации уже известных методов их анализа

Эффективность решения задачи обработки текстов зависит от решения проблемы автоматизации их анализа. С другой стороны, автоматизация анализа текста подчиняется глобальным практическим задачам, связанным с обнаружением механизма построения текста и выявлением его характерных свойств.

Использование математических методов при решении подобных задач обеспечивает получение объективных результатов, расширяет число применяемых методов и приемов при исследовании текстов, а также дает возможность решения таких задач, сама постановка которых без применения данных методов может быть нереальной.

Одной из наиболее актуальных задач квантитативного языкознания является необходимость создания теории, которая позволила бы описать и объяснить закономерности организации связного текста.

В последние годы были получены новые результаты, выявившие некоторые закономерности построения текста, построены математические модели этой организации. Большую роль в этих исследованиях сыграли работы Ю.К. Орлова, Ю.А. Шрейдера, М.В. Арапова, Е Н Ефимовой, Б.В. Сухотина, Б.И. Кудрина, Ю.К. Крылова и др , которые показали, что в природе существует закон, который управляет механизмом формирования структуры текстов. Но механизм формирования этой структуры до конца не выявлен.

Широкое распространение получил статистический метод анализа структуры текста, который, в частности, сводится к оценке рангового закона распределения Устойчивость таких законов распределения как гиперболического (H – распределения) и закона Ципфа по отношению к объектам различной природы является свидетельством о наличии в природе закона, который управляет механизмом формирования структуры текста.

Классическая теория вероятностей не исследует причины формирования определенного вида закона распределения, а закон распределения не раскрывает онтологической природы текста и является только внешним проявлением пока неизвестной его внутренней структуры Поэтому произошла смена парадигмы - изменение представления о природе и свойствах текста Согласно новой парадигме текст рассматривается как некоторая целостность, но в литературе отсутствуют какие-либо результаты исследований в этом направлении.

Поэтому, предлагаемый в диссертации метод построения модели структурной организации текста и разработка соответствующего алгоритма его обработки является актуальным.

Особую благодарность автор выражает профессору Нижегородского государственного лингвистического университета Глебовой Е.Ф., советы которой помогли в работе над диссертацией.

Цель работы. Целью работы является построение модели структурирования текста и разработка алгоритма для его обработки, а также разработка методики идентификации текстов

Задачи работы. Достижение намеченной цели требует решения следующих задач:

- Построение модели структурирования текста как некоторой целостности и ее сравнение с уже существующими моделями;

- Разработка алгоритма статистической обработки текста с целью выяснения его структуры;
- Описание структуры текста с помощью информационной матрицы, далее в диссертации названной "информационным портретом";
- Разработка методов идентификации текстов на основе сравнения "информационных портретов" текстов.

Методы исследований.

Методологической основой данной работы является системный анализ. В качестве математического аппарата использованы элементы теории вероятностей и математической статистики, элементы теории информации.

Научная новизна.

- Разработана новая модель структурирования текста как некоторой целостности, достоверность которой подтверждается в результате анализа рангового закона распределения;
- Введено новое информационное описание структурных связей между языковыми единицами в тексте;
- Разработана методика идентификации текстов на основе сравнения "информационных портретов";

Обоснованность и достоверность результатов обеспечена доказательствами сформулированных в работе положений и представленными результатами статистической обработки текстов

Практическая ценность заключается в возможности применения предложенной модели как нового инструмента при анализе структуры текста в целях поиска информации, исключении ошибок при переводе и передаче информации, а также возможности идентификации текстов и построении частотных словарей в целях изучения лексики текстов, написанных на различных языках.

Реализация результатов работы.

Разработанные в рамках диссертационной работы алгоритм анализа структуры текста и методика идентификации текстов на основе сравнения "информационных портретов" используются в учебном процессе Нижегородского государственного лингвистического университета им. Н.А Добролюбова.

Апробация результатов работы

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

- VI-ой Международной конференции "Математика. Компьютер. Образование" (Пущино, 1999);
- Международной конференции "Математика Образование. Экология Гендерные проблемы" (Воронеж, 2000),
- Международной конференции "НТИ-2000. Информационные технологии и телекоммуникации" (Москва, ВИНИТИ, 2000),
- Научно-технической конференции факультета Информационных систем и технологий ФИСТ-2000 (Н.Новгород, НГТУ, 2000);
- Всероссийской научно-технической конференции "Информационные системы и технологии ИСТ-2001 (Н. Новгород, 2001).
- Всероссийском научно-практическом семинаре "Проблемы прикладной лингвистики" (Пенза, 2001).

Публикации.

По теме диссертационного исследования опубликовано 10 работ

Структура и объем диссертации.

Диссертация состоит из введения, трех глав, заключения, списка литературы и приложений Общий объем работы 132 страницы текста, содержащего 26 рисунков и 9 таблиц Список литературы содержит 127 наименований.

СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Во введении дается общая характеристика работы, обосновываются актуальность темы диссертации, цели теоретических и практических исследований, показана научная новизна и практическая ценность диссертационной работы. Аннотировано по главам излагается содержание диссертации.

В первой главе приведен обзор литературы по имеющимся методам исследования структурной организации текста.

Показано, что многообразие существующих направлений и моделей при анализе структуры текста отражает практические потребности в разработке адекватного и эффективного аппарата анализа текста, позволяющего усовершенствовать действующие и разработать новые системы автоматической обработки информации. Поставлена задача разработки модели структурирования текста.

Во второй главе описан новый дедуктивный метод построения модели структурирования текста, приводятся результаты статистической обработки текстов с целью оценки вероятностей появления в них языковых единиц.

В диссертации рассматриваются ранговые законы распределения вероятностей, отмечается, что зависимость Ципфа-Мандельброта позволяет получить приближенное описание статистической организации текста, но оставляет в стороне вопрос о причинах, ответственных за структуру этой организации.

В качестве объекта исследования текст выступает как структурированная целостность многомерный семантически организованный объект. Семантические корреляции распространяются на весь текст в целом, его развертывание является процессом специфической природы. Поэтому построение теории самоорганизации текста как

открытой системы с атрибутом целостности требует построения новых моделей его структуры.

Рассмотрим текст, понимаемый как список вхождения словоформ. Каждой словоформе соответствует некоторое слово. Совокупность всех слов, образующих текст **T** назовем словарем **V** данного текста. Для каждого слова **W** из словаря **V** укажем целое число **n(W)**, равное количеству имеющихся в тексте словоформ, которым соответствует данное слово **W**. Величину **n(W)** назовем встречаемостью слова **W** в тексте **T**.

Общая сумма встречаемости слов будет равна количеству словоформ **N** в тексте **T** или объему этого текста. Упорядочим теперь слова в словаре по убыванию **n(W)**. Номер слова в таком списке назовем рангом **K**, а слово ранга **K** обозначим W_k. Ранги слов будут принимать значения от 1 до M, где **M** – общее число слов в словаре. Тогда относительную частоту W_k / N можно использовать в качестве оценки вероятности P_k появления слова ранга **K** в тексте.

Для этих вероятностей Ципфом эмпирически был установлен закон распределения:

$$P_k = \frac{C}{K^{\gamma}},$$

где γ - некоторая постоянная.

Распределения, подобные распределению Ципфа известны не только в лингвистике. Они существуют в биологии, экономике, социологии, науковедении и др.

В диссертации используется дедуктивный метод построения внутренней структуры текста, закономерности которой обнаруживаются в модифицированном ранговом законе распределения. Этот метод основан на применении закона "золотого деления" к тексту в целом, согласно

этому закону целое так относится к большей части, как большая часть к меньшей.

Последовательное развертывание этого метода приводит к возможности выдвижения гипотезы, которая сводится к утверждению, что среди всех словарных единиц текста (букв алфавита) существуют словарные единицы, вероятность появления которых в тексте равна

$$\left(0,5 * (\sqrt{5} - 1)\right)^{k} \approx 0,618^{k}.$$

где 0,618 –значение "золотого деления", а k некоторое целое число, свое для каждой словарной единицы (буквы).

Эти значения всроятностей образуют базис, в котором определяются вероятности остальных словарных единиц как линейная форма. Если по оси ординат откладывать логарифм вероятности, а по оси абсцисс – ранг словарной единицы, т.е. ее порядковый номер в последовательности из словарных единиц, расположенных в порядке убывания их вероятностей. то получается прямая линия, при этом предполагается, что словарные единицы с равными вероятностями имеют одинаковые значения рангов и на графике представлены одной точкой.

Приводятся результаты статистической обработки текстов, отмечается высокая точность совпадения оценки вероятностей с их априорными значениями, вычисленными по предлагаемой методике, что подтверждает адекватность разработанной модели реальной структуре текста, в то время как закон Ципфа не всегда выполняется, в частности, он не описывает закон распределения букв.

Если описать данные структуры разработанной моделью. и вместо логарифма ранга по оси абсцисс отложить значения рангов, а по оси ординат логарифм вероятности, то фактические данные (рис 1,2, точки) с высокой точностью совпадают с теоретической прямой, обозначенной сплошной линией (рис. 1,2).

Например, различные буквы алфавита лежат на нескольких прямых, параллельных главной прямой и находящихся на некотором расстоянии от нее (выше и ниже) (рис 1)

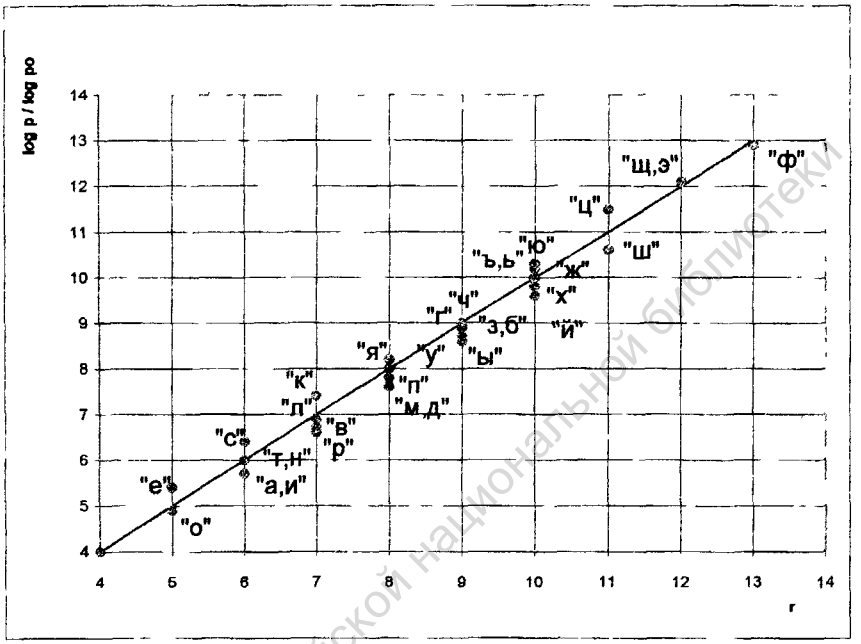

Рис.1 Модифицированный ранговый закон распределения букв русского алфавита

Аналогичные результаты были получены при обработке текста произведения А С. Пушкина "Капитанская дочка". Ранее исследование данного текста проводилось в работах Ю К. Орлова и Р М Фрумкиной, где было показано, что описание частотной структуры текста подчиняется закону Ципфа-Мандельброта (закону Ципфа). Однако следует констатировать, что на графиках, приведенных в этих работах, видны существенные отклонения от линии, соответствующей точному выполнению закона Ципфа.

Если предположить, что структура текста произведения А.С Пушкина "Капитанская дочка" описывается предлагаемой в настоящей работе моделью и вместо логарифма ранга отложить по оси абсцисс значения рангов и логарифм вероятности по оси ординат, то фактические данные с более высокой точностью приближаются к теоретической прямой (рис. 2).

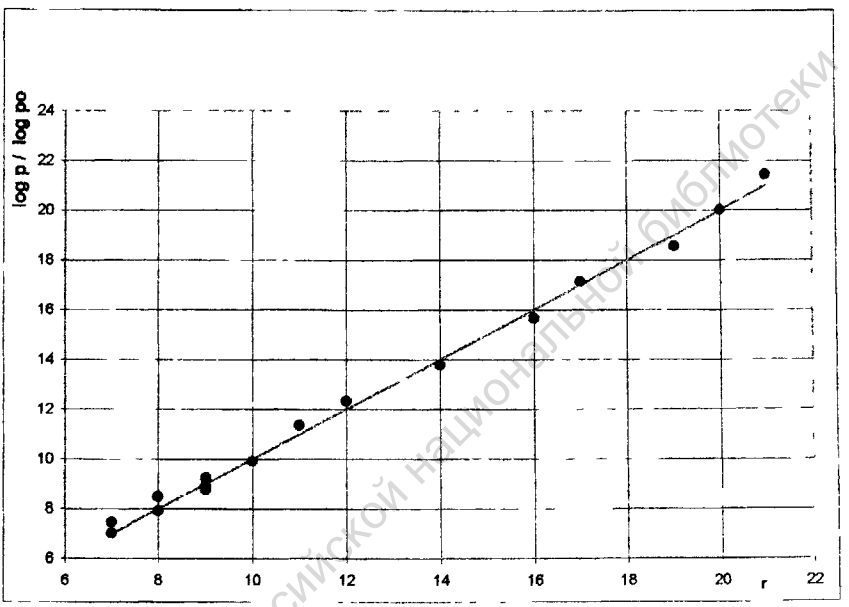

Рис.2. Модифицированный ранговый закон распределения слов из произведения А С. Пушкина «Капитанская дочка»

Таким образом, истинность модели структурирования текста подтверждается в результате анализа рангового закона распределения текста. Данная модель справедлива для распределений, не подчиняющихся закону Ципфа (распределению букв русского алфавита) и распределений, которые приближенно описываются законом Ципфа – в текстах литературных произведений и частотных словарях

В третьей главе диссертационной работы разработана методика идентификации текста. Предлагаемый алгоритм статистической обработки

текста позволяет построить своеобразный "портрет" его структуры и осуществить идентификацию и классификацию соответствующих "портретов".

В отличие от известных методов статистической обработки текста, в работе предлагается описывать статистическую зависимость между словами (классами слов) не только посредством условных вероятностей, но и с помощью взаимной информации.

"Информационный портрет" структуры текста строится на множестве всевозможных комбинаций из двух слов (классов слов), находящихся на заданном расстоянии друг от друга в предложении.

Каждую комбинацию из двух классов слов можно изобразить точкой в декартовой системе координат Таким образом, каждой комбинации из двух классов слов ставится в соответствие количественная мера взаимной информации между ними·

$$I(x_i, y_j) = \log\left(p(x_i, y_j) / p(x_i)p(y_j)\right)$$

где $p(x_i,y_j)$ - вероятность появления пары классов слов x_i и y_j, i и j - порядковые номера классов слов на координатных осях; $p(x_i)$ и $p(y_j)$ - безусловные вероятности появления классов слов в тексте.

При построении "информационного портрета" текста ожидаемые результаты исследования связаны с фиксацией формальных различий текстов, принадлежащих разным авторам, и сходстве текстов одного и того же автора. Апробация данной методики проводилась на текстах А.С. Пушкина, А.П. Чехова, А. Куприна.

В этих текстах был проведен грамматический анализ слов – разбиение всего множества слов на 8 следующих классов· имя существительное, глагол, имя прилагательное, местоимение, наречие, числительное, причастие и класс служебных частей речи Затем текст каждой выборки был переведен в последовательность кодов. Обработка

закодированного текста данных выборок позволила оценить для каждой из них:

1) матрицу **B** вероятностей парных встречаемостей классов слов

$$B=\begin{bmatrix} p(x_1,y_1) & p(x_1,y_2) & \dots & p(x_1,y_n) \\ p(x_2,y_1) & p(x_2,y_2) & \dots & p(x_2,y_n) \\ \dots & \dots & \dots & \dots \\ p(x_n,y_1) & p(x_n,y_2) & \dots & p(x_n,y_n) \end{bmatrix}$$

где $p(x_i,y_j)$ - вероятность появления пары классов слов x_i и y_j, i и j - номера классов слов ($i=1,..\ n; j=1,..\ n; n=8$); $p(x_i)$, $p(y_j)$ - безусловные вероятности появления классов слов в тексте.

2) все возможные связи между классами слов можно изобразить в виде матрицы **A** - строками и столбцами которой являются классы слов, а на пересечении строки и столбца находится взаимная информация между этими классами:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}, \quad (n=8)$$

где

$$a_{ij} = \log(p(x_i,y_j) / p(x_i)p(y_j))$$

матрица A названа "информационным портретом" Сравнение структур текстов производится посредством сравнения соответствующих "информационных портретов"

Для сравнения структур текстов были использованы коэффициент корреляции **K** и среднеквадратическое отклонение σ^2 :

$$K = \frac{K_1}{S_1 S_2}$$

где:

$$K_1 = \frac{\sum_{i,j} a^1_{ij} a^2_{ij}}{m},$$

$$S_1 = \sqrt{\frac{\sum_{i,j} \left(a^1_{ij}\right)^2}{m}}, \quad S_2 = \sqrt{\frac{\sum_{i,j} \left(a^2_{ij}\right)^2}{m}}$$

a^1_{ij} – элементы информационной матрицы первого текста;

a^2_{ij} – элементы информационной матрицы второго текста.

$$\sigma^2 = \frac{\sum_{i,j} \left(a^1_{ij} - a^2_{ij}\right)^2}{m}$$

m – число пар в сравниваемых текстах.

Величина коэффициента корреляции ($K \approx 0{,}93$) при сравнении текстов одного и того же автора оказалась больше, чем при сравнении текстов разных авторов ($K \approx 0{,}8$) (табл. 1).

Таблица 1

Значение величины коэффициента корреляции при сравнении текстов одного автора и разных авторов

автор / автор	А.С Пушкин	**А.П. Чехов**	А. Куприн
А С. Пушкин	K=0,91373	K=0,80488	K=0,85858
А.П. Чехов	K=0,80488	K=0,94107	K =0,79046
А. Куприн	K=0,85858	K =0,79046	K=0,91068

Это свидетельствуст о влиянии на структуру текста индивидуальных особенностей автора. Наличие сильной корреляции между любыми текстами объясняется тем, что в текстах проявляется одна и та же структура языка, при этом каждый автор задает свой функциональный механизм его использования (рис. 3).

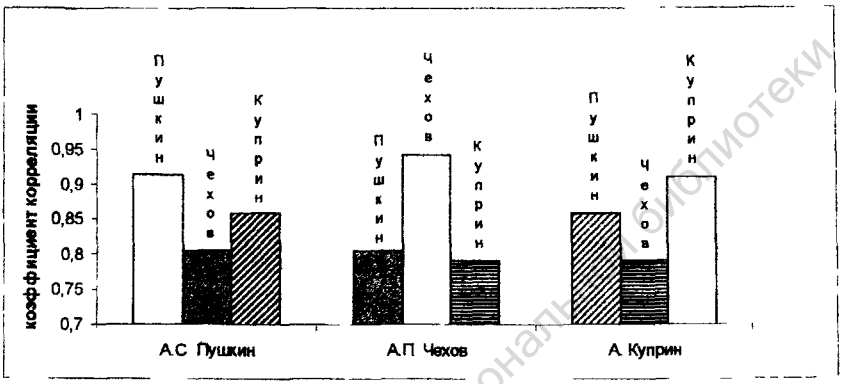

Рис.3 Сравнительный анализ текстов по величине коэффициента корреляции

Аналогичные результаты получены при сравнении матриц по величине среднеквадратического отклонения σ^2 ($\sigma^2 \approx 0{,}07$ для текстов, принадлежащих одному и тому же автору; $\sigma^2 \approx 0{,}17$ для текстов, принадлежащих разным авторам) (табл.2).

Таблица 2

Значение величины среднеквадратического отклонения при сравнении текстов одного автора и разных авторов

автор / автор	А.С. Пушкин	А.П. Чехов	А. Куприн
А.С. Пушкин	σ^2 =0,083351	σ^2 = 0,16503	σ^2 =0,15874
А.П. Чехов	σ^2 =0,16503	σ^2 =0,06658	σ^2 =0,19661
А. Куприн	σ^2 =0,15874	σ^2 =0,19661	σ^2 =0,09067

Различия, наблюдаемые при сравнении "информационных портретов" текстов, принадлежащих одному и тому же автору, связаны с погрешностями статистической обработки и с изменением структуры индивидуального языка автора во времени. При сравнении "портретов" текстов, принадлежащих разным авторам, наблюдается увеличение величины среднеквадратического отклонения, связанное с различиями в индивидуальных стилях авторов (рис.4)

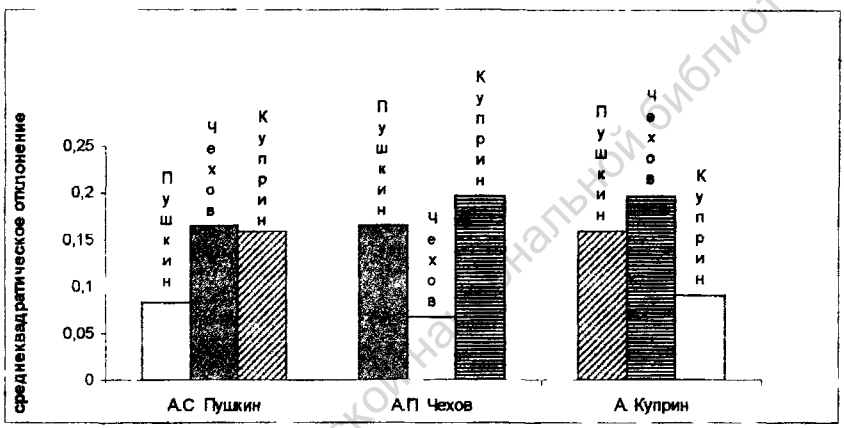

Рис.4. Сравнительный анализ текстов по величине среднеквадратического отклонения

На основании анализа "информационных портретов" текстов разных авторов и текстов одного автора можно сделать вывод о том, что различия между "портретами" произведений одного автора значительно меньше различий, наблюдаемых при сравнении "портретов" произведений разных авторов

Таким образом, та пара текстов, которая характеризуется наибольшим коэффициентом корреляции $\mathbf{K}$ и наименьшим среднеквадратическим отклонением $\boldsymbol{\sigma}^2$ наиболее вероятно принадлежит одному автору.

Приложения содержат результаты статистической обработки текстов с целью получения рангового закона распределения и сравнительного анализа текстов, документы, подтверждающие использование и внедрение результатов работы в учебном процессе.

Основные результаты работы

- Разработана новая модель структурирования текста, достоверность которой подтверждается в результате анализа рангового закона распределения.
- Введено новое описание структурных связей с помощью взаимной информации.
- Разработан алгоритм статистической обработки текста, позволяющий построить "информационный портрет" его структуры.
- Разработана методика идентификации текстов на основе сравнения "информационных портретов"
- Проведена идентификация текстов различных авторов по разработанной методике.

Список публикаций по теме диссертации:

1 Ломакина Л.С., Панкратова А.З Анализ некоторых моделей лингвистических явлений // Математика. Компьютер. Образование. Вып.6 Часть I. Сборник научных трудов. Под ред. Г Ю Ризниченко М Прогресс-Традиция, 1999. С 102-105

2. Панкратова А.З. Сетевое моделирование как метод исследования некоторых лингвистических явлений // Исследования молодых ученых: Сборник статей аспирантов. Часть III. Мн МГЛУ, 1999 С.59-61.

3. Глебова Е.Ф., Ломакина Л С , Панкратова А З Моделирование сложного синтаксического целого как структурно-семантического

единства // Математика Образование. Экология. Гендерные проблемы. Материалы международной конференции. Том 1. Воронеж: НОУ «Интерлингва», 2000. С.213.

4. Ломакина Л.С., Панкратова А.З. Оптимизационные методы лингвистической дешифровки // Системы обработки информации и управления. Межвузовский сборник. Вып.6. Н. Новгород, 2000 С.74-78.

5. Ломакин Д.В , Ломакина Л.С., Панкратова А.З. Вероятностно-информационная модель для исследования структуры текста// Научно-техническая конференция факультета информационных систем и технологий. ФИСТ-2000. Н. Новгород· НГТУ, 2000. С.113-114.

6. Глебова Е.Ф., Ломакина Л С., Ломакин Д.В, Панкратова А.З. Информационно-статистическая модель в лингвистических исследованиях. // Пятая международная конференция "НТИ-2000. Информационные ресурсы и технологии. Телекоммуникации". М.: ВИНИТИ, 2000. С.83-84

7. Ломакин Д.В., Панкратова А.З. Модель структурирования текста // Всероссийская научно-техническая конференция, посвященная 65-летию факультета информационных систем и технологий. ФИСТ-2001. Н. Новгород: НГТУ, 2001.С.177.

8. Панкратова А.З Построение модели структуры текста // В творческом поиске. Сб научных трудов аспирантов. Ч.2. Мн.. МГЛУ, 2001. С.26-34.

9 Ломакин Д.В., Панкратова А.З. Модель структуры текста// Системы обработки информации и управления. Межвузовский сборник. Вып.7. Н. Новгород, 2001. С.99-103.

10. Панкратова А.З. Идентификация текста на основе информационной модели сго структуры // Всероссийский научно-практический семинар "Проблемы прикладной лингвистики" Пенза: Приволжский дом знаний, 2001. С.40-41.

Лицензия ЛР № 020073 от20.06.97

Подписано к печати 26.04.2002 Формат 60x84 1/16

Печ. л. 1,1. Тираж 100 экз.

Заказ

Типография НГЛУ им. Н.А. Добролюбова

603155, Нижний Новгород, ул. Минина 31а