

УДК 343.98

**МАТЕМАТИЧЕСКИЕ МЕТОДЫ КРИМИНАЛИСТИЧЕСКОЙ ДИАГНОСТИКИ И АВТОРИЗАЦИИ ТЕКСТА В РЕЧЕВЕДЧЕСКОЙ ЭКСПЕРТИЗЕ**

© 2016 г.

**В.А. Юматов,<sup>1</sup> М.В. Маркина,<sup>1</sup> С.В. Юматов<sup>2</sup>**<sup>1</sup>Нижегородский государственный университет им. Н.И. Лобачевского, Н. Новгород<sup>2</sup>Главное управление

Министерства внутренних дел Российской Федерации по Нижегородской области, Н. Новгород

yumatovva@yandex.ru

*Поступила в редакцию 01.09.2016*

Проанализированы отдельные общенаучные методы судебно-экспертных исследований речеведческих объектов с использованием компьютерных технологий, где на основе авторской программы анализируются свойства личности, дается система признаков и характеристика их информативности.

*Ключевые слова:* компьютерные технологии, методы экспертного исследования, программно-аппаратные комплексы, алгоритмы нахождения слов текста и знака, подсчета и вычисления индекса Флеша и Ганнинга в исследуемом тексте.

Одним из действенных показателей науки считается использование ею математических и кибернетических методов исследования, относящихся ко второй и третьей группе общенаучных методов судебно-экспертных исследований. При использовании этих методов объекты и процессы (явление, событие) рассматриваются как сложные целостные системы, а их исследование ведется с позиций системного подхода. Для понимания процессов и явлений, общих законов получения, хранения, передачи и преобразования информации в сложных управляющих системах (кибернетический метод) нужно уметь выделять необходимые параметры, определяющие функционирование системы, и выявлять взаимосвязи между ними.

Независимо от того, что анализируется, операция анализа с логической точки зрения осуществляется одинаково, это расчленение целого на части, то есть системы на составляющие ее элементы. Синтезирующее исследование состоит в интеграции отдельных элементов. Это метод систематизации, в результате применения которого получают новое знание. Эти два метода также взаимосвязаны. Синтез немислим без анализа, т.е. без операции расчленения и выявления отдельных элементов целого, представляющего объект исследования. Системное целостное представление об объекте возможно, если объект исследования показан в виде систем, а системные задачи могут представлять системный анализ (определение свойств системы по известной ее структуре) и (или) системный синтез (определение структуры системы по ее свойствам). Следует отметить, что объект экспертного исследования «рассматривается

как сложная динамическая система, состоящая из трех элементов: 1) материальный носитель информации о данном факте, событии; 2) источник информации о факте; 3) механизм передачи информации от источника к носителю, другими словами, отражаемый и отражающий компоненты и механизм их взаимодействия» [1].

Судебные эксперты и те, кто организует эту деятельность (самоуправляемая система), являются участниками и одновременно свидетелями изменения роли судебной экспертизы в системе правоприменения в условиях современного развития технологий, подходов и вызовов глобального мира. Экспертные технологии в настоящее время поднялись на высокий уровень, а их использование позволяет в значительной мере автоматизировать рутинные операции при производстве судебных экспертиз, повысить точность и оперативность сложных и громоздких расчетов и т.п. Данные тенденции характерны для каждого класса судебных экспертиз, в том числе и экспертиз речеведческих. Отметим, что в настоящее время можно выделить одно из наиболее изученных направлений применения современных компьютерных технологий, – автоматизацию процесса сбора и обработки экспериментальных данных, получаемых при выполнении фоноскопических экспертиз с помощью специальных приборов, объединенных с компьютером. Например, можно выделить такие системы, как «Диалект», «ОТExpert», «Justiphone», «ИКАР Лаб»: они позволяют производить идентификацию личности по фонограммам устной речи; устанавливать аутентичность (достоверность) аналоговых и цифровых фонограмм речи; анализировать шумы и

акустическую обстановку и условия проведения звукозаписи; осуществлять идентификацию средств звукозаписи; проводить повышение качества и разборчивости фонограмм речи и др. [2]. Эти технологии значительно расширяют возможности судебных экспертов при решении конкретных вопросов, так как вводятся критерии оценки качества экспертных исследований, выполняемых с использованием различных измерений.

Не пытаясь на современном, явно недостаточном для этого материале построить определенную систему знаний, попробуем несколько пополнить фрагментарное представление о свойствах личности и закономерностях, определяющих зависимости от них речеведческих объектов. В настоящем исследовании в целях получения интересующих нас данных об исполнителе текстов обосновываются принципы (физичности, моделируемости и целенаправленности) и предлагается модель комплексного исследования, позволяющего аккумулировать имеющуюся на момент их построения совокупность разрозненных отрывков информации об искомом (расследуемом) событии.

Авторами статьи демонстрируется только начальный модуль технологии анализа и интерпретации информации, содержащейся в материалах, предоставляемых в распоряжение судебного эксперта. Сущность предложенного метода выражается в выявлении и оценке имеющихся в тексте лингвистических признаков разных уровней (текстового, синтаксического, лексического, морфологического, фонетического), которые характеризуют различные компоненты текста (денотативный, оценочный, иллокутивный, экстралингвистический), то есть предписывают причинно-следственные связи объекта данной природы (принцип физичности). Заданные параметры могут быть расширены в зависимости от целей исследователя, а также созданы новые модули, отражающие определённую грань сущности, тем самым обеспечивается возможность использования в системном подходе упрощенных моделей (принцип моделируемости). Для реализации этих целей используется программный комплекс обработки текстовой информации с учетом особенностей русского языка. Согласно предлагаемому подходу, сначала исследуемый текст, предназначенный для обработки, должен быть подготовлен в виде текстового файла или введен в онлайн-режиме пользователем программы. Для проверки правильности написания слов применяется словарь, который является текстовым файлом. Инструкция по работе с программой загружается при выборе пункта меню «Справка».

В программе реализованы следующие алгоритмы:

1. Алгоритм формирования массива слов текста и знаков препинания после каждого слова. Список возможных знаков препинания задается.

2. Алгоритм определения частоты встречаемости слов и знаков препинания.

3. Алгоритм нахождения наиболее употребляемых слов и знаков препинания.

4. Алгоритм нахождения прилагательных в тексте и подсчета их количества. Прилагательные идентифицируются по окончанию слова.

5. Алгоритм нахождения глаголов в тексте и подсчета их количества. Глаголы идентифицируются по окончанию слова.

6. Алгоритм формирования массива предложений, из которых состоит исходный текст.

7. Алгоритм подсчета среднего количества слов в предложениях текста.

8. Алгоритм поиска слов в словаре, использующий бинарный поиск элемента в массиве. Для поиска замены ошибочно написанного в тексте слова используется понятие «дистанция Левенштейна». Эта «дистанция» – минимальное количество правок первого слова (под правками подразумеваются три возможные операции: стирание символа, замена символа и вставка символа), чтобы превратить его во второе слово.

9. Алгоритм вычисления индекса Флеша – Кинкейда (FRE-индекс).

Индекс Флеша – это мера оценки сложности текста. Вычисляется по формуле:

$$FRE = 206.835 - (1.3 \times ASL) - (60.1 \times ASW),$$

где:

ASL – средняя длина предложения в словах (англ. Average Sentence Length);

ASW – средняя длина слова в слогах (англ. Average Number of Syllables per Word).

Индекс по шкале FRES (Flesch Reading Ease Scale) распределяется таким образом:

100 и выше: очень легко читается. Средняя длина предложения составляет 12 или менее слов. Нет слов более чем из двух слогов;

от 65 до 100: простой язык. Средняя длина предложения составляет от 15 до 20 слов. В среднем слова имеют 2 слога;

от 65 до 30: труднее читается. Предложения содержат до 25 слов. Обычно двусложные слова;

от 0 до 30: очень трудно читать. В среднем предложение имеет 37 слов. Слово имеет в среднем более 2 слогов.

10. Алгоритм вычисления индекса Роберта Ганнинга (FOG-индекс).

Индекс Ганнинга показывает, какой образовательный уровень нужен для усвоения данного материала. Чем индекс меньше, тем большей аудитории он будет понятен.

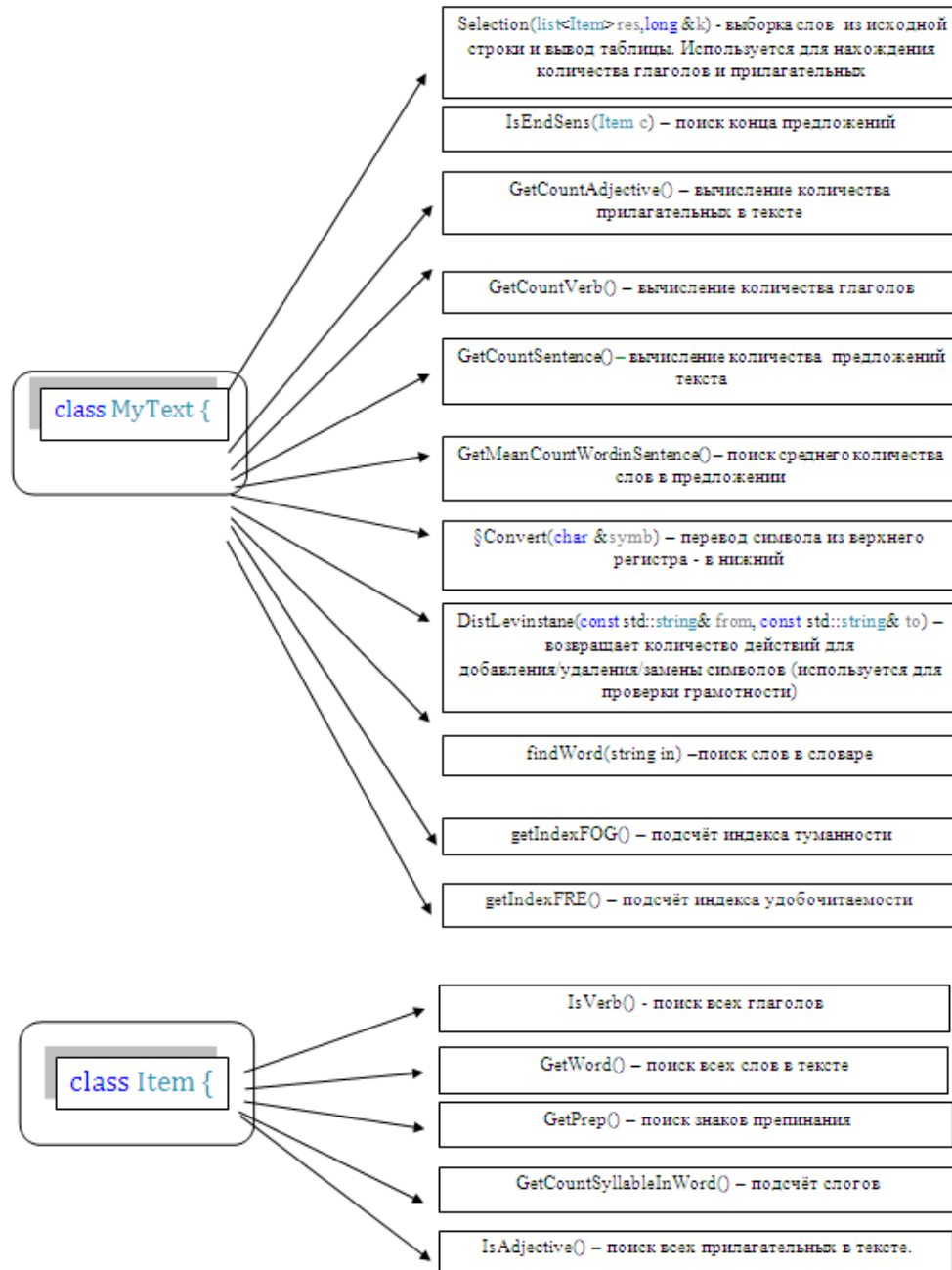


Схема программы с описанием основных функций

Для русскоязычных текстов вычисляется по формуле

$$\text{Индекс Ганнинга} = 0.4 \left[ 0.78 \left( \frac{\text{слов}}{\text{предложений}} \right) + 100 \left( \frac{\text{число сложных слов}}{\text{число слов}} \right) \right],$$

где: число сложных слов – количество слов с числом слогов больше трёх; 0.78 – поправочный коэффициент для русского языка.

Считается, что индекс в пределах от 16 до 20 подходит для людей с высшим образованием; 13–15 – для студентов младших курсов;

9–12 – уровень старших классов (или газетный уровень); 8 и ниже – уровень школьников.

Программа написана на языке C++ в среде Visual Studio 2012.

Рассмотрим пример обработки текстовой информации с помощью данной программы.

Считанный из файла текст – один из федеральных законов (рис. 1).

На рис. 2 показаны результаты обработки текста с помощью представленной программы. На рис. 2, табл. 1, помещены результаты деления текста на слова, с определением знаков препинания после каждого слова и идентификацией прилагательных и глаголов. На рис. 2,

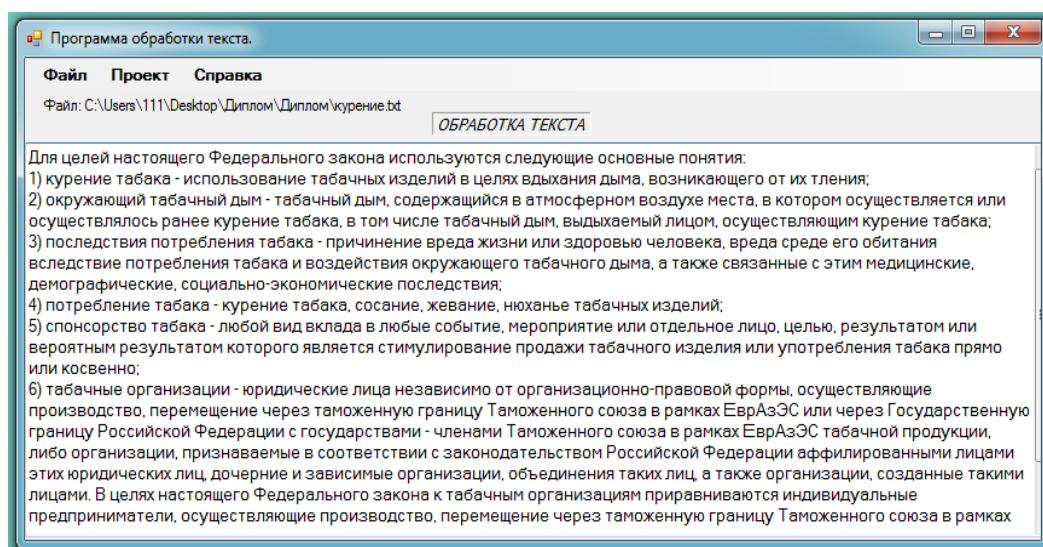


Рис. 1

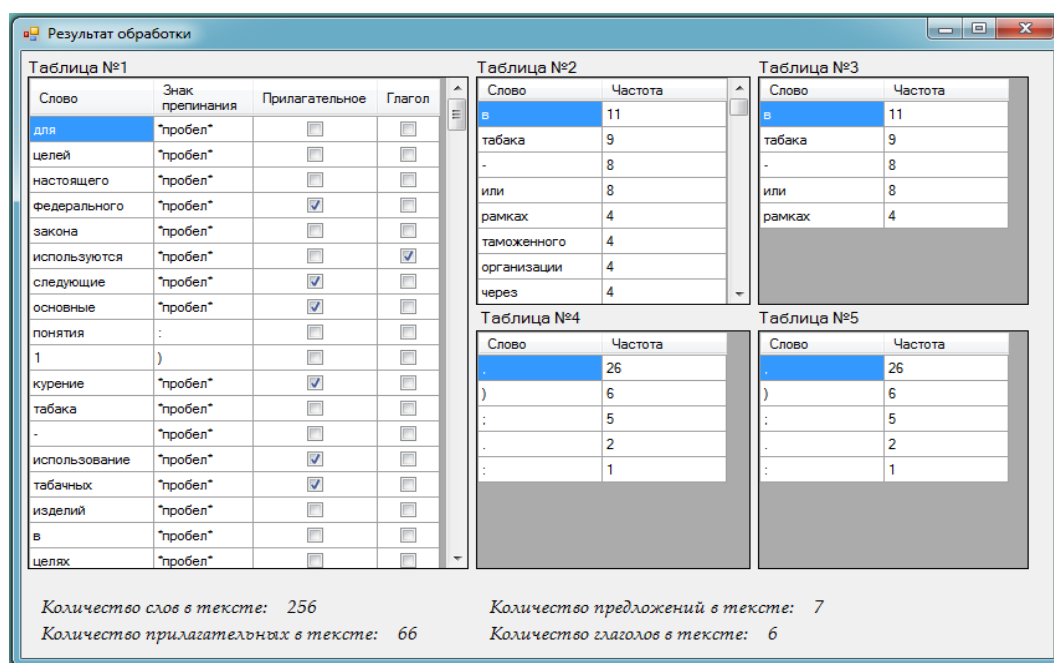


Рис. 2

табл. 2, – результат подсчета частоты встречаемости для каждого слова текста. На рис. 2, табл. 3, представлены самые популярные слова из обработанного текста. На рис. 2, табл. 4, – результат подсчета частоты встречаемости каждого знака препинания. На рис. 2, табл. 5, помещены самые популярные в этом тексте знаки препинания. Также выдается информация о том, сколько в тексте слов, предложений, прилагательных, глаголов.

Рис. 3 – результат вычислений индексов Флеша и Ганнинга для данного текста.

Результаты проверки другого произвольно введенного текста на грамматические ошибки представлены на рис. 4 и 5.

Кратко сформулируем промежуточные итоги исследования. Предложенный метод экспертно-

го исследования ориентирован на решение конкретных задач:

- во-первых, на выявление классификационно значимых характеристик текста (языковые, социально-биографические, психологические, ситуативные);
- во-вторых, выявление функциональных стилей языка (научного, официально-делового, газетно-публицистического), разговорной и ораторской речи, специфики художественных текстов;
- в-третьих, выражение смыслов в исследуемых текстах и определение формы их выражения; установления объема и содержания понятия, выражаемого словом или словосочетанием, употребленным в тексте.

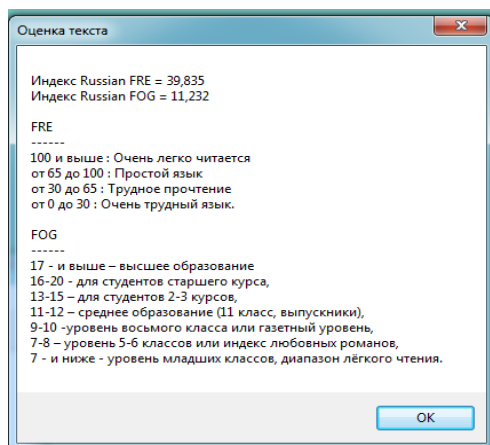


Рис. 3

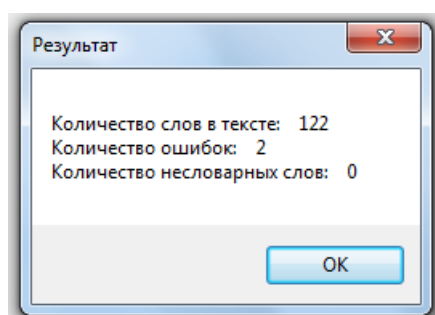


Рис. 4

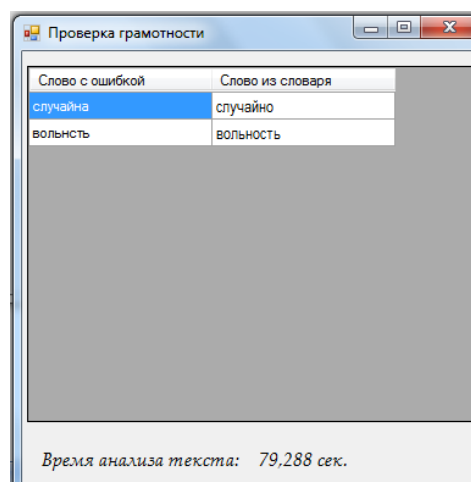


Рис. 5

Именно поэтому предлагаемый программный продукт криминалистической диагностики и авторизации текста будет способствовать повышению эффективности обработки русскоязычных текстов за счет синтеза качественных и количественных характеристик и, следовательно, может быть использован в учебном процессе при подготовке судебных экспертов по специальности 40.05.03 «Судебная экспертиза», а также судебных экспертов, специализирующихся на производстве речеведческих экспертиз. Расширение сферы использования математических и кибернетических методов исследования, алгоритмизации и информатизации экспертной деятельности будет способ-

ствовать модернизации существующих и созданию принципиально новых баз знаний и экспертных систем [3].

#### Список литературы

1. Россинская Е.Р., Галяшина Е.И., Зинин А.М. Теория судебной экспертизы. М.: Норма, ИНФРА-М, 2011. С. 382.
2. Бурцева Е.В., Селезнёв А.В., Чернышов В.Н. Информационные технологии в юриспруденции: Учеб. пособие. Тамбов: Изд-во Тамб. гос. техн. ун-та, 2012. 104 с.
3. Малков В.П., Маркина М.В. Поэтапная параметрическая оптимизация: Учебное пособие. Н. Новгород, 1998. 142 с.

## MATHEMATICAL METHODS FOR FORENSIC DIAGNOSTICS AND AUTHOR IDENTIFICATION IN LINGUISTIC EXAMINATION

*V.A. Yumatov, M.V. Markina, S.V. Yumatov*

In this article, we present an analysis of some scientific methods for forensic examination of text objects using computer technology and describe a linguistic text processing program developed by the authors. The program analyzes the properties of an individual based on text elements and features. A system of features is presented and information content of these features is characterized.

*Keywords:* computer technologies, methods of forensic examination, hardware and software systems, algorithms for finding words and punctuation marks, calculation of Flesch Reading Ease Score and Gunning Fog Index in the text.

*References*

1. Rossinskaya E.R., Galyashina E.I., Zinin A.M. Teoriya sudebnoj ehkspertizy. M.: Norma, INFRA-M, 2011. S. 382.
2. Burceva E.V., Seleznyov A.V., Chernyshov V.N. Informacionnye tekhnologii v yurisprudencii: Ucheb. posobie. Tambov: Izd-vo Tamb. gos. tekhn. un-ta, 2012. 104 s.
3. Malkov V.P., Markina M.V. Poetapnaya parametricheskaya optimizaciya: Uchebnoe posobie. N. Novgorod, 1998. 142 s.