

Федеральное государственное бюджетное образовательное  
учреждение высшего образования  
«Томский государственный университет систем управления и  
радиоэлектроники» (ТУСУР)

На правах рукописи



Исхакова Анастасия Олеговна

**МЕТОД И ПРОГРАММНОЕ СРЕДСТВО  
ОПРЕДЕЛЕНИЯ ИСКУССТВЕННО СОЗДАНЫХ ТЕКСТОВ**

05.13.17 – Теоретические основы информатики

Диссертация на соискание ученой степени кандидата технических наук

Научный руководитель  
доктор технических наук,  
профессор Р.В. Мещеряков

Томск – 2016

## Содержание

Введение.....	4
1 Существующие методы и алгоритмы определения происхождения текста	13
1.1 Методы текстовой атрибуции.....	13
1.1.1 Методы атрибуции, основанные на статистическом анализе .	16
1.1.2 Методы атрибуции, основанные на использовании искусственного интеллекта.....	21
1.2 Автоматическая генерация текстов.....	24
1.2.1 Метод генерации текстов на основе использования цепи Маркова.....	24
1.2.2 Метод генерации текстов на основе SIMP-таблиц.....	28
1.2.3 Метод генерации текста с использованием словарей .....	30
1.2.4 Метод генерации текста, основанный на синонимизации .....	32
1.3 Методы определения массово порожденных текстов.....	36
1.4 Алгоритм определения искусственно созданных текстов.....	39
1.5 Выводы.....	41
2 Разработанные инварианты искусственно созданных текстов .....	43
2.1 Классификация текстовых произведений на основе инвариантов ...	43
2.2 Предложенный алгоритм формирования инвариантов классов текстов .....	46
2.3 Формирование инвариантов искусственных и естественных текстов на основе предложенного алгоритма .....	54
2.3.1 Определение качественных признаков выделенных классов текстов.....	56
2.3.2 Определение свойств текста, определяющих проявление качественных признаков .....	57
2.3.3 Формирование множества количественных характеристик ...	58

2.3.4 Расчет значений характеристик текста для наборов текстов известного происхождения .....	62
2.3.5 Проверка перечня характеристик на различительную способность .....	64
2.3.6 Оценка взаимозависимости характеристик .....	65
2.3.7 Полученные инварианты искусственных текстов.....	66
2.4 Выводы.....	68
3 Метод и программное средство определения искусственно созданных текстов .....	69
3.1 Предлагаемый метод определения искусственно созданных текстов.....	69
3.2 Программное средство фильтрации искусственно созданных текстов.....	75
3.2.1 Требования к программному средству .....	76
3.2.2 Структура программного средства .....	77
3.3 Тестирование метода и программного средства.....	87
3.4 Выводы.....	92
Заключение .....	94
Список использованной литературы.....	97
Приложения .....	114

## ВВЕДЕНИЕ

**Актуальность вопроса.** Использование текстовой формы представления данных для хранения и передачи сведений повсеместно используется людьми во всех сферах жизни. Создание текста долгое время было связано исключительно с умственной деятельностью человека – его автора. На сегодняшний день тексты имеют не только функцию хранения, накопления и передачи информации. Благодаря развитию способов моментального обмена, а также возможности распространения данных посредством сети Интернет, авторы текстов имеют широкие возможности донести информацию до большого количества читателей.

Объемы создаваемой текстовой информации за последние десятилетия постоянно возрастают, об этом свидетельствует развитие дата-центров, интернет-ресурсов различного назначения, электронного документооборота и т.д. Вместе с тем создание самих текстов уже не является уникальной прерогативой человека. Специальные алгоритмы и программные средства позволяют генерировать тексты автоматически на основе некоторых исходных данных. **В дальнейшем в диссертационной работе тексты, созданные автоматически с помощью специальных алгоритмов или программных генераторов, будут называться искусственно созданными или искусственными.**

Методы создания искусственных текстов позволяют генерировать множество уникальных экземпляров на основе некоторого авторского текста или на основе описанной модели формирования текста. Они широко применяются для создания информационного контента, так как размещаемая в сети Интернет информация должна обладать достаточным уровнем уникальности, чтобы сайт был проиндексирован поисковыми системами и пользователи смогли его найти.

Интернет и различные инфокоммуникационные технологии с каждым годом занимают все большее место в общественных отношениях и

взаимодействиях на всех уровнях. В этой связи информация, распространяемая в сети, зачастую воспринимается человеком как современный аналог энциклопедии или справочника, а также телевидения и газет, которым люди привыкли верить. Однако есть основания полагать, что эта информация не всегда отражает действительность и может быть использована для введения пользователей Интернета в заблуждение, распространения заведомо ложной или подстрекательской информации и в других неправомерных целях.

Массовое автоматическое порождение текстов определенной направленности может быть нацелено на пропагандирование различных идей, в том числе социального, политического, а также преступного характера, манипуляцию населением или парализацию работы и выведение из строя определенных электронных ресурсов [1]. Использование искусственных текстов в виде информационного контента популярных или специально разработанных для этого веб-ресурсов позволяет публиковать и распространять уникальные тексты любого содержания в неограниченном количестве. Таким образом, учитывая неоспоримую значимость Интернет-технологий в жизни человека, злоумышленники могут использовать их в собственных неправомерных целях.

Задача определения искусственных текстов напрямую связана с исследованиями текстов для получения сведений об условиях их создания, объединенными текстовой атрибуцией. Методическое обеспечение в данной области начало развиваться с исследованиями по определению авторства, однако на сегодняшний день текстовая атрибуция включает в себя более широкий спектр задач. Проблема анализа и выявления искусственных текстов пока не получила достаточной освещенности, и на сегодняшний день не разработано действенных методов для выявления такого класса текстов. В то же время распространение искусственно созданной текстовой информации только возрастает.

Указанные угрозы для общества и то, что на сегодняшний день задача автоматического выявления такого контента не решена, обуславливают необходимость создания метода определения искусственно созданных искусственных текстов.

**Состояние вопроса.** Определение свойств текста, в том числе связанных с его происхождением, неразрывно связано с методами, используемыми в текстовой атрибуции в целом. Исследование характеристик текста с целью определения, кем и когда он был написан, в первую очередь связано с задачами определения авторства. Определение авторства на сегодняшний день является активно изучаемой и расширяющейся областью научных знаний. Российскими и зарубежными учеными проводится множество исследований для оценки существующих и разработки новых методов.

Среди ученых, достигших значительных результатов в определении авторства текстовых произведений, можно выделить Д.В. Хмелева [2, 3], который в своих исследованиях доказал и подробно описал возможность использования статистического анализа значения текстовых характеристик в решении задач текстовой атрибуции, в частности при определении авторства. Им также разработано программное средство, позволяющее определять авторов по рассчитанным для них образам, которые также называют инвариантами.

Наибольшей точности в определении авторства – до 95 % верных решений – удалось добиться в своих работах:

– К. Кирееву – предложенная методика основана на статистическом анализе ряда характеристик текста, а также текстовых штампов и реализована в виде программного средства «Штампомер» [4];

– О.Г. Шевелеву – в программном средстве «Стилеанализатор» [5, 6] использован смешанный подход: использованы статистический анализ, деревья решений, нейронные сети;

– А.С. Романову – предложенная им методика и разработанное программное средство «Авторовед» основаны на использовании аппарата опорных векторов [7].

В последние пять лет внимание ученых в области определения авторства было сконцентрировано на определении авторства коротких текстов: статей в Интернете, сообщений электронной почты, sms-сообщений, записей в социальных сетях и т.п. Наибольшее распространение получило определение авторства сообщений электронной почты. В этом направлении достигнуты значительные успехи, что обусловлено достаточной длиной подобных текстов. Результаты таких исследований опубликованы, например, в работе австралийского ученого М. Corney с соавторами [8], где приведены результаты изучения электронных сообщений на основе 184 параметров (использовались 253 текста четырех авторов на английском языке). В работе указана средняя точность идентификации авторства – 82%.

Также работы по определению авторства сообщений электронной почты опубликованы индийским ученым Т. Chakraborty [9], американскими авторами К. Calix, М. Connors и др. [10]. Изучение авторства микросообщений на примере Твиттера проводила группа ученых из Канады – М.Л. Brocardo, I. Traore и др. [11]. Авторство sms-сообщений также исследовано учеными из США R. Zheng, J. Li и др. [12], R.G. Ragel, P. Herath и др. [13].

В последние три года в связи с распространением и доступностью программных генераторов текста, а также возрастающим информационным воздействием, исходящим из сети Интернет, стали актуальными исследования текстов на предмет их происхождения, а именно установления способа создания.

Среди перспективных исследований неестественных текстов также можно выделить исследования А.С. Павлова, Б.В. Доброва с соавторами по разработке методов обнаружения поискового спама [14, 15]. Кроме того

известны работы, формализующие подходы к исследованию веб-спама, авторства таких ученых, как С. Castillo, D. Donato [16], а также Z. Gyöngyi, H. Garcia-Molina [17, 18]. Особенности естественных текстов, отраженные в работе А.В. Анисимова с соавторами [19], также косвенно относятся к решаемой задаче.

В исследованиях текстов, созданных автоматически, кроме символьных и лексических характеристик уделяют внимание тематическим и семантическим. Для оценки текстов в таком случае часто используются методы автоматической рубрикации текстов, подробно описанные в работах Б.В. Доброва с соавторами [20, 21], Е.В. Дунаева и А.А. Шелестова [22], в которых особое внимание уделяется рассмотрению веб-страниц и особенностям работы с ними.

Несмотря на очевидный интерес мирового сообщества к исследованиям искусственно созданных текстов, на сегодняшний день не разработан универсальный метод определения текстов, сгенерированных программными алгоритмами, которая бы с высокой точностью (> 90% верных решений) определяла такие тексты.

**Целью диссертационной работы** является повышение точности определения искусственно созданных текстов за счет создания авторского метода и программного средства.

**Объектом исследования** являются простые текстовые сообщения, сгенерированные автоматически на основе специальных алгоритмов и исходные естественные тексты.

**Предметом исследования** являются методы и алгоритмы классификации текстов.

Для достижения указанной цели необходимо решение следующих задач:



1) исследовать методы и алгоритмы создания инвариантов, оценить возможность их применения для создания инвариантов текстов, сгенерированных автоматически;

2) выделить инварианты для классификации естественных и искусственно созданных текстов;

3) разработать метод определения искусственно созданных текстов;

4) провести экспериментальную апробацию метода определения искусственно созданных текстов;

5) разработать программный комплекс, позволяющий определять искусственно созданные тексты.

**Методы исследования.** Для решения поставленных задач в диссертационной работе использовались методы функционального и математического моделирования, теории множеств, математической статистики, матричных вычислений.

**Научная новизна** проведенных исследований и полученных в работе результатов заключается в следующем:

1. Создан оригинальный метод определения искусственно созданных текстов, отличающийся использованием меры принадлежности входного текста к известным классам текстов и позволяющий принять решение о происхождении текста.

2. Разработан новый алгоритм формирования инвариантов классов текстов, отличающийся от существующих использованием качественных и уточняющих их количественных текстовых характеристик и позволяющий осуществить выбор компонентов инварианта на основе лингвистических особенностей текстов.

3. Предложены новые инварианты для текстов, созданных искусственно с помощью синонимизации и метода Марковских цепей, полученные с использованием авторского алгоритма и позволяющие провести классификацию текстов по способу их создания.

**Практическая значимость результатов работы.** Разработанный метод и основанное на нем программное средство позволяют решить задачи автоматической фильтрации интернет-контента, входящих сообщений и иных электронных текстов. Использование разработанного программного средства позволяет определить нежелательные страницы, содержащие искусственно созданный контент, оценить приходящие онлайн-запросы, а также идентифицировать потенциально опасные, вредоносные текстовые сообщения на электронных ресурсах.

**Положения, выносимые на защиту:**

1. Метод определения искусственно созданных текстов, основанный на расчете меры принадлежности входного текста к известным классам, позволяет определить, является исследуемый текст естественным или искусственным в 93 % случаев. Соответствует п. 6 паспорта специальности 05.13.17.

2. Алгоритм формирования инвариантов классов текстов описывает последовательность действий для определения обладающих различительной способностью качественных, а также уточняющих их количественных характеристик текста, значения которых формируются в инварианты. Соответствует п. 5 паспорта специальности 05.13.17.

3. Инварианты естественных текстов и текстов, созданных искусственно с помощью синонимизации и метода цепей Маркова, позволяют представить различия данных классов текстов формально – в виде набора значений характеристик и использовать их при определении искусственных текстов. Соответствует п. 2 паспорта специальности 05.13.17.

4. Программное средство фильтрации искусственно созданных текстов позволяет производить автоматическое определение способа создания входного текста. Соответствует п. 9 паспорта специальности 05.13.17.

**Достоверность результатов** обеспечивается проверкой непротиворечивости и адекватности результатов, полученных как на промежуточных, так и на окончательных этапах работы.

**Внедрение результатов работы.** Результаты диссертационной работы были внедрены в деятельность ООО «Агентство медиарешений» и ООО «Лингвистические и информационные технологии», а также в учебную деятельность ТУСУРа по дисциплинам «Дискретная математика», «Теория вероятностей и математическая статистика».

**Личный вклад.** В диссертационной работе использованы результаты, в которых автору принадлежит определяющая роль. Часть опубликованных работ написана в соавторстве с научным руководителем.

Диссертант является автором метода определения искусственно созданных текстов, представленного в работе, а также разработчиком комплекса программ. Программное обеспечение для расчета значений характеристик текста зарегистрировано Федеральной службой интеллектуальной собственности.

Автором совместно с коллективами предприятий ООО «Агентство медиарешений» и ООО «Лингвистические и информационные технологии» проведены внедрение и апробация результатов работы. Постановка задачи исследования осуществлялась научным руководителем доктором технических наук, профессором Р.В. Мещеряковым.

**Апробация работы.** Основные результаты диссертационной работы докладывались на следующих семинарах и конференциях:

1) Томских IEEE-семинарах «Интеллектуальные системы моделирования, проектирования и управления», г. Томск, 2013–2016 гг.;

2) XIV Всероссийской конференции молодых ученых «Актуальные проблемы лингвистики и литературоведения», г. Томск, 2013 г.;

3) XIII Всероссийском конкурсе-конференции студентов и аспирантов по информационной безопасности «SIBINFO-2013», г. Томск, 2013 г.;

4) XVIII Всероссийской научно-технической конференции студентов, аспирантов и молодых учёных «Научная сессия ТУСУР-2013», г. Томск, 2013 г.;

5) Межвузовской научно-практической конференции «Актуальные проблемы инфосферы. Инфокоммуникации. Геоинформационные технологии. Информационная безопасность», г. Санкт-Петербург, 2013 г.;

6) Шестой международной конференции по когнитивной науке, г. Калининград, 2014 г.;

7) II Всероссийской научной интернет-конференции с международным участием «Современные системы искусственного интеллекта и их приложения в науке», г. Казань, 2014 г.;

8) 11-ой международной научно-технической конференции «Интерактивные системы: проблемы человеко-компьютерного взаимодействия», г. Ульяновск, 2015 г.;

9) Седьмой международной конференции по когнитивной науке, г. Светлогорск, 2016 г.

Результаты диссертационной работы использовались при выполнении проекта «Методы и алгоритмы идентификации моделей поведения объектов информационных инфраструктур для обеспечения безопасности государства, общества и личности», поддержанного грантом РФФИ № 16-47-700350.

**Публикации по теме диссертации.** Результаты диссертационной работы отражены в 11 публикациях, в том числе 5 публикаций в рецензируемых журналах из перечня ВАК, 6 публикаций в сборниках трудов конференций

**Объем и структура работы.** Диссертация состоит из введения, трех глав, заключения. Полный объем диссертации составляет 123 страницы с 12 рисунками, 8 таблицами. Список использованных источников содержит 121 позицию.

# 1 СУЩЕСТВУЮЩИЕ МЕТОДЫ И АЛГОРИТМЫ ОПРЕДЕЛЕНИЯ ПРОИСХОЖДЕНИЯ ТЕКСТА

## 1.1 Методы текстовой атрибуции

Первые исследования, посвященные разработке и применению методов формального определения авторства текстов, относятся к началу XX в. Однако споры об авторстве различных произведений возникали задолго до этого. Предпосылками подобных споров являлись случаи плагиата публикации анонимных текстов, творчество под псевдонимами и т.д.

В мировой практике известно множество громких дел, связанных с выдающимися произведениями всемирно признанных авторов [23–25]. Например, сомнению подвергалось авторство таких произведений, как «Слово о полку Игореве», «Тихий Дон», «Мертвые души». К этому списку можно добавить не один десяток художественных и исторических произведений разных эпох [26]. Несмотря на большое количество исследований, проведенных для разрешения таких споров, окончательного и точного ответа по ним на сегодняшний день нет. Это связано с тем, что достоверная проверка полученных результатов практически невозможна или затруднена, а уличение в плагиате и неподлинности текстов всемирно признанных авторов должно быть обосновано. Тем не менее результаты большинства вышеуказанных исследований свидетельствуют об эффективности методов обработки текстового материала по формальным признакам.

В исследованиях по определению авторства долгое время использовались субъективные методы, например, оценка внешних деталей авторского стиля (употребление «любимых» слов, терминов, выражений).

При этом в первой половине XX в. доминировали филологические и историко-документальные методы исследования текстов, однако середина прошлого столетия характеризуется появлением работ по определению авторства, в которых применялись математико-статистические методы [27, 28]. Развитие в конце 1970-х гг. методов автоматизации обработки текстов и расчета их численных характеристик ознаменовало новый этап в этом научном направлении. Резкое повышение вычислительных возможностей за счет применения ЭВМ [29] позволило ученым добиться значительного повышения эффективности в определении авторства текстовых произведений.

Задача определения происхождения текста неразрывно связана с необходимостью выявления характерных языковых особенностей и стилистических приемов, используемых в произведении. Нередко для решения такой задачи привлекаются эксперты, которые могут идентифицировать автора неизвестного текста либо определить факт принадлежности текста другому автору. Поскольку определение происхождения текста востребовано в различных сферах деятельности, а экспертный анализ авторского стиля является трудоемким и длительным процессом, актуальным является создание и развитие формальных методов решения подобных задач.

Под текстовой атрибуцией [30] понимается исследование текста для получения сведений об условиях его создания. Среди задач текстовой атрибуции можно выделить следующие [31]:

- определение авторства [32] (получение сведений о лице, написавшем текст);
- диагностический анализ текста [33–35] (получение сведений о личностных характеристиках автора текста, таких как пол, возраст, темперамент, уровень образования и т.д.);

– верификация авторства [36] (получение ответа на вопрос, создан ли текст некоторым автором или группой авторов).

Выделенные задачи направлены на разрешение литературоведческих вопросов, проведение социальных исследований, выявление фактов нарушения авторского права, криминалистических экспертиз и расследований [37].

В последнее десятилетие перечень задач, связанных с атрибуцией текстов, постоянно расширяется. Это обусловлено в первую очередь изменениями в способах создания, распространения и обработки текстового материала, а также его предназначения. Развитие сети Интернет и многократное повышение скорости обмена информацией в современном мире обусловили проблему распространения массового порожденных текстов.

На сегодняшний день среди множества работ ученых всего мира не представлено общепризнанного и эффективного решения задачи определения искусственных текстов. Кроме того, рассматриваемая в данном исследовании задача однозначно не классифицирована.

С одной стороны, она имеет схожие черты с задачей верификации авторства, так как требуется определить, имеет ли некоторый программный генератор отношение к созданию данного текста. С другой стороны, определение искусственно созданных текстов можно отнести к задаче идентификации авторства, так как в данном случае необходимо определить, каким из известных способов создан текст, либо сделать вывод о том, что это определить невозможно.

Автором предпринята попытка проанализировать применение существующих методов текстовой атрибуции для решения задачи определения искусственно созданных текстов.

Современные методы атрибуции текстов основываются на статистическом анализе либо машинном обучении (применении

искусственного интеллекта) [38]. Такие методы, как правило, позволяют исследовать текстовое произведение на пунктуационном, орфографическом, синтаксическом, стилистическом и лексико-фразеологическом уровнях [39]. Это позволяет отразить основные свойства текста, выделяемые специалистами в области лингвистики для большинства языков – символные, лексические, синтаксические, семантические и тематические, – связанные с советующими текстообразующими компонентами: знаками, словами и словосочетаниями (лексемами), предложениями, абзацами и т.д.

### **1.1.1 Методы атрибуции, основанные на статистическом анализе**

В основе класса методов, основанных на статистическом анализе, лежит тот факт, что стиль автора можно определить по значению определенного параметра или набора параметров – авторскому инварианту. Примерами таких параметров могут быть средняя длина слова, частота вхождений некоторого символа, набора символов или определенных слов.

Под авторским инвариантом понимается количественная характеристика литературных текстов, поведение которой однозначно описывает произведения одного автора или небольшого числа «близких авторов» и которая принимает существенно разные значения для произведений разных групп авторов [40]. Чаще всего используется набор численных значений характеристик текста.

Применение статистического анализа в решении задачи атрибуции текстов основано на следующих этапах:

а) выделяется инвариант и определяется критическая граница для некоторого авторства;

б) текущее измеренное значение инварианта для текста с неизвестным авторством сравнивается с ранее определенным критическим граничным значением;



в) в зависимости от положения измеренного значения на числовой оси относительного критической границы принимается решение о том, что текст с высокой вероятностью принадлежит автору либо, напротив, с высокой вероятностью не принадлежит автору.

Статистические методы нашли широкое применение в вопросах атрибуции. К их достоинствам по сравнению с машинными методами можно отнести быстрое проведение расчетов, возможность обучения, универсальность [41].

Недостатком таких методов является необходимость выделения инварианта, что требует дополнительных статистических расчетов. Объем расчетов зависит не только от задачи, стоящей перед исследователем, но и от имеющейся выборки текстовых произведений и ее свойств.

В то же время малейшие ошибки в определении инварианта могут значительно повлиять на результат работы алгоритма. При этом под ошибкой понимается использование текстовых характеристик, которые по тем или иным причинам не дают возможности различить авторов.

В работе Фоменко А.Т. [23], посвященной исследованию больших объемов текста, были выделены следующие требования к характеристикам текста, используемым для составления инвариантов при решении задач, связанных с текстовой атрибуцией:

а) массовость – возможность измерения характеристики для любого текстового произведения, под массовостью также понимается слабая контролируемость автором на сознательном уровне;

б) устойчивость – некоторое постоянное среднее значение характеристики для одного автора или группы авторов, слабое отклонение значений от среднего;

в) различительная способность – свидетельство того, что для разных авторов или групп авторов выделенная характеристика текста принимает

разное значение. При этом измеренное в каждом случае значение позволяет отнести текст лишь к одному из авторов (классов).

Свойства характеристик обладать массовостью, устойчивостью и различительной способностью не зависят от языка, на котором написаны тексты. Исследования показывают, что для разных языков стилеобразующие характеристики повторяются, но применяются согласно нормам соответствующего языка [42].

Примером могут служить частота употребления слов определенных частей речи, употребление  $n$ -грамм [43] символов. Такие характеристики часто относятся к лексическим, морфологическим, структурным, контекстно-специфическим. Синтаксические и идиосинкразические характеристики в большей степени зависят от грамматических особенностей и не имеют закономерных зависимостей между разными языками. Исключением могут быть языки, схожие по культуре и истории возникновения, однако этот факт требует дополнительной проработки в каждом конкретном случае.

Перечисленные свойства приняты исследователями как условие применимости формальных параметров в методиках атрибуции текста. Зачастую при решении некоторой прикладной задачи налагаются дополнительные требования к используемым характеристикам, возникающие из-за наложенных ограничений (например, упомянутая выше мультязычность).

Статистический анализ используется в исследованиях, посвященных текстовой атрибуции в целом и определению авторства в частности. Он применяется во многих опубликованных методиках определения авторства и реализованных на их основе программных средствах [44]. Предложенные методики различаются уровнем наполненности набора выделенных в авторские инварианты характеристик, которые чаще всего не распространяются в открытом виде, а вынесены на обозрение только частично или скрыты.

Несмотря на то, что многие из проведенных исследований в области текстовой атрибуции показали положительные результаты, для входных текстов применяются существенные ограничения. Причинами этого являются недостаточная база данных авторов, исследованных при создании системы, и различия в объеме входных данных, необходимых для обработки [45].

Невозможность использования большинства существующих систем обусловлена также и тем, что некоторые из них созданы для решения узких задач, связанных с проведением литературоведческих исследований: доказательства или опровержения причастности конкретного человека к определенному литературному произведению.

В диссертационной работе были рассмотрены различные методики и программные средства, выделены их особенности, в том числе в области применимости на различном текстовом материале. В качестве критериев в обзоре были выбраны следующие:

- используемые характеристики текста;
- возможность обучения, то есть использования методики или программного средства для решения собственной задачи;
- преимущества решения, позволяющие использовать его для решения задачи определения искусственных текстов;
- накладываемые ограничения.

Методикам определения авторства на основе статистического анализа на сегодняшний день удается добиться до 98 % точности в решении задачи. Средний показатель составляет 70–85 %. Среди программных средств по определению авторства, основанных на авторских методиках с использованием средств статистического анализа, можно выделить следующие:

- 1) Программа «Штампомер» [4], основанная на автоматической статистической оценке текстовых параметров: знаков препинания,

количества слов в предложении, повторения штампов (словосочетаний) от 1 до 5 слов. «Штампомер» не предполагает возможность расширения набора исследуемых характеристик или дополнительного обучения. Минимальный объем текста для обработки должен составлять 30 тыс. символов.

2) «Лингвоанализатор» [46] – программное средство, разработанное Д.В. Хмелевым на основе авторской методики. В его основе сравнение текста с существующими наборами авторских инвариантов, которые внесены в программу с помощью статистического анализа. Программа не поддерживает возможность обучения. Для идентификации необходимо не менее 10 тыс. символов.

3) Программа «Атрибутор» [47] – разработка О.В. Замолуева, А.Н. Тимашева и А.А. Поликарпова, позволяющая определить авторство русских текстов 103 писателей. В основе методики, на которой построена работа программы, лежит статистический анализ, механизм цепей Маркова. Для обработки текста необходимо ввести не менее 20 тыс. символов. Программа не подразумевает возможность обучения.

4) Лингвистический анализатор [48], созданный группой ученых и программистов, работа которого построена на статистическом анализе – оценке отклонений от средних значений характеристик текста. 10 тыс. символов – минимальный объем обрабатываемого текста. Существует возможность обучения, так как авторами предлагается открытый код программы.

Обзор данных решений по определению авторства показал следующее:

1) большинство решений можно отнести к решению узких прикладных задач, таких как определение авторства для текстов определенного литературного периода или определенного перечня авторов (100–150 авторов), для определения автора короткого текста или текста определенной структуры (электронное письмо, заявка, sms-сообщение);

2) все решения активно используют в своей работе значения символьных и лексических характеристик текста как слабоконтролируемых автором на сознательном уровне и позволяющих оценить словарный запас автора;

3) рассмотренные программные средства и методики не подразумевают возможность обучения для исследования других групп текстов;

4) для достоверного определения авторства необходимы тексты объемом от 10 Кб, в некоторых программах введена специальная проверка вводимого текста по данному критерию;

5) в связи с многообразием текстовых произведений высокой эффективностью определения авторства считаются методики и программные средства на их основе, позволяющие достоверно классифицировать тексты в более 70% испытаний.

### **1.1.2 Методы атрибуции, основанные на использовании искусственного интеллекта**

Искусственным интеллектом называются технологии, которые позволяют машине (компьютеру) выполнять умственные, исследовательские, творческие функции, которые традиционно выполняются человеком или при участии человека [49]. Чаще всего они связаны с разработкой интеллектуальных программ или машин, которые включают в себя базы знаний в некоторой предметной области. Механизмы искусственного интеллекта отличает способность создавать в ходе самообучения эвристические программы для решения задач определенного класса сложности и решать эти задачи. Возможность данных механизмов работать со слабоструктурированными данными позволяет решать задачи, связанные с

анализом показателей, зависимости между которыми не определены в достаточной степени или даже неизвестны [50, 51].

Отличием методов атрибуции, основанных на использовании механизмов искусственного интеллекта, является то, что процесс формирования инвариата частично или полностью возлагается на машину, они являются логическим развитием методов математической статистики, переведенных на более высокий уровень вычислений. Мнение эксперта при использовании данных методов заменяется результатами работы систем принятия решений. Если при статистическом анализе эксперт отбирает ряд характеристик, составляющих инвариант, то искусственный интеллект позволяет выделить инвариант автора, отличающий его от остальных, автоматически после обучения классификатора [52].

Среди наиболее распространенных методов определения авторства, основанных на интеллектуальном анализе данных [53], можно выделить искусственные нейронные сети, машину опорных векторов, генетические алгоритмы, деревья решений и др.

В последнее десятилетие выросла популярность использования нейронных сетей в различных областях, в том числе в задачах классификации, к которым относится и задача определения автора текста. Нейронные сети – эффективный метод моделирования, позволяющий воспроизводить сложные зависимости. Они нелинейны по своей природе и инвариантны к размерности данных, что позволяет моделировать зависимости в случае большого числа переменных [54]. Для обучения нейронной сети необходимы обучающая выборка данных и набор входных параметров. Алгоритм обучения сети автоматически воспринимает структуру данных.

Нейронные сети являются мощным инструментом в решении задач прикладной статистики, однако использование данного метода требует

значительных вычислительных затрат, а также наличия достаточной обучающей и тестовой выборок.

Методы, основанные на обучении нейронных сетей, а также на машине опорных векторов, во многом благодаря своей адаптивности и достаточной простоте обучения классификатора показали хорошие результаты в задачах определения авторства в ряде предложенных учеными методик: точность в определении автора текста варьируется от 65 до 98%. Однако выработанные с помощью средств искусственного интеллекта правила не всегда очевидны для человека, так как такие механизмы чаще всего работают по принципу «черного ящика».

Среди методов и программных решений для определения авторства текста, основанных на использовании механизмов искусственного интеллекта, можно выделить следующие:

1) методику и программное средство определения авторства «Авторовед» [7], работа которой основана на методе опорных векторов и использовании нейронных сетей; точность работы данного решения достигает 98% при тексте длиной от 20 тыс. символов и выше;

2) программу «JGAAP» [55, 56], основанную на применении Байесовского классификатора и метода главных компонент, которая является кроссплатформенным решением для определения авторов, однако данное решение не поддерживает ряд языков, в том числе русский.

3) программное средство «Стилеанализатор» [5, 6] является примером смешанного подхода к определению авторства: в нем использованы статистический анализ, деревья решений, нейронные сети [57].

Следует отметить, что, несмотря на эффективность машинных классификаторов, решения в области текстовой атрибуции на практике не обходятся без статистических расчетов для формирования инвариантов и проведения экспериментов, подтверждающих практическую работоспособность и достоверность реализуемого метода.

## 1.2 Автоматическая генерация текстов

С распространением искусственных, то есть сгенерированных автоматически, текстов развиваются и методы их создания. Это связано в первую очередь с тем, что искусственно созданный контент, не обладающий достаточной уникальностью, блокируется поисковыми системами [58], а они на сегодняшний день являются основным проводником между пользователем Интернета и веб-ресурсами. В зависимости от длины и назначения создаваемого текста также могут использоваться различные методы генерации [59]. В данном исследовании будут рассмотрены генераторы, на основе которых может быть создан информационный контент для интернет-ресурса.

Уникальность генерируемых текстов оценивалась посредством алгоритма шинглов [60, 61], который активно используется поисковыми системами для поиска нечетких дубликатов веб-страниц. Данный алгоритм основан на сравнении контрольных сумм последовательностей лексем нескольких текстов. Данные последовательности называются шинглами и могут иметь различную длину. Поиск нечетких дубликатов, реализуемый методом шинглов, позволяет рассчитать степень схожести двух объектов. Под объектами в данном случае понимаются текстовые файлы.

### 1.2.1 Метод генерации текстов на основе использования цепи Маркова

В практике автоматической генерации текстов наиболее распространенным методом является использование простой цепи Маркова. Цепью Маркова называется последовательность испытаний, в каждом из которых появляется только одно из  $k$  несовместных событий  $A_i$  из полной группы. При этом условная вероятность  $p_{ij}(s)$  того, что в  $s$ -м испытании



наступит событие  $A_j$  при условии, что в  $(s - 1)$ -м испытании наступило событие  $A_i$ , не зависит от результатов предшествующих испытаний [62]. Популярность данного метода генерации текстов обусловлена возможностью создавать уникальные тексты, обладающие низким количеством грамматических ошибок и несогласованностей, а также простотой программной реализации.

Применительно к генерации текста метод генерации текстов на основе использования цепи Маркова можно трактовать следующим образом:

- а) исходный текст разбивается на лексические единицы;
- б) для каждой уникальной лексической единицы составляется массив, состоящий из самой лексемы и слов, которые могут располагаться после нее (на основе исходного материала);
- б) выбирается начальная лексема и помещается в текст-результат;
- в) среди вероятных слов, которые могут следовать за выбранной лексемой, выбирается единственное и помещается в текст-результат;
- г) выбранная лексема является новым звеном, для которого вновь выбирается следующее за ним слово.

Алгоритм может быть ограничен количеством символов, лексических единиц или предложений. Опытные исследования показывают, что максимально человекоподобный текст получается из больших объемов исходного текста – более 20 тыс. символов. Также отмечается, что текст должен быть единого тематического содержания, то есть не содержать в себе разные предметные области [63]. Например, исходный текст с упоминанием медицинских терминов и подробностей устройства автомобиля повлечет некорректный с точки зрения восприятия текст-результат.

Наглядный пример генерации текста на основе представления лексических единиц в виде цепи Маркова, представленный в [64], приведен ниже. Он показывает принцип формирования текста на основе исходного.

Схожесть текста-результата с исходным, измеренная алгоритмом шинглов [65], составляет 46,67% для шингла из 1 слова и 23,53% для шингла из 2 слов. В источнике отмечается, что низкая уникальность текста-результата связана с очень малым размером исходного текста (19 слов).

Исходный текст:

*Из-за леса, из-за гор едет дедушка Егор:*

*сам на лошадке, жена на коровке,*

*дети на телятках, внуки на козлятках.*

Разбиение на лексические единицы и вероятные связки (в квадратных скобках):

1. из-за [леса, гор]
2. леса [из-за]
3. гор [едет]
4. едет [дедушка]
5. дедушка [Егор]
6. Егор [сам]
7. сам [на]
8. на [лошадке, коровке, телятках, козлятках]
9. лошадке [жена]
10. жена [на]
11. коровке [дети]
12. дети [на]
13. телятках [внуки]
14. внуки [на]

Текст-результат: *Егор сам на телятках внуки на лошадке жена на коровке дети на коровке.*

Пример изменения текста, сгенерированного на основе текста объемом 1 830 слов (11 500 символов), повествующего о событиях Куликовской битвы, приводится ниже.

Фрагмент исходного текста:

*«8 сентября 1380 года возле впадения в Дон реки Непрядвы произошло сражение, получившее название Куликовской битвы. Итогом боя на Куликовом поле стала полная победа русских. О мифах и исторических фактах великого события читайте в материале».*

Сгенерированный текст-результат:

*«На поле, расположенном возле впадения в Дон реки Непрядвы, супротив друг друга на рассвете выстроились русские и татарские войска. Едва рассеялся туман, началась сеча, продолжавшаяся как минимум три часа. Итогом ее стала полная победа русских полков, которые еще долго гнали бежавшего противника по степи. История сия известна каждому мало-мальски грамотному россиянину, ее изучают в школе, она кочует из учебника в учебник уже не одно столетие. Естественно, этими сухими фактами дело не ограничивается, и рассказ о ходе битвы, о погибших князьях и боярах и другие подробности».*

Как видно на примере, текст-результат доступен для чтения и понимания смысла, в нем практически нет грамматических ошибок, однако находится множество повторов словосочетаний длиной 3 и более лексемы. Это связано с наличием редких слов, географических наименований, редко используемых просторечий в исходном тексте, а также его малым объемом.

Схожесть текста-результата с исходным (полным вариантом), измеренная алгоритмом шинглов, составляет 31,12% для шингла из 1 слова и 22,43% для шингла из 2 слов. То есть при использовании исходного материала менее 2 000 слов возможно достичь уровня уникальности текста-результата свыше 60%.

### 1.2.2 Метод генерации текстов на основе SIMP-таблиц

SIMP-таблицы (с англ. – Simplified Integrated Modular Prose – упрощенная интегрированная модульная проза) – таблицы с некоторыми фрагментами предложений, например, таблицы *A*, *B*, *C*, *D*, в которых записаны начала предложений, внутренние части и завершения. Из каждой такой таблицы выбирается случайное значение, и таким образом формируется законченное предложение [61, 66]. Возможны варианты изменения порядка таблиц, например *B*, *A*, *C*, *D*, а также добавление механизмов для коррекции окончаний слов при необходимости.

Данный метод позволяет генерировать общеупотребительные псевдонаучные фразы, как, например, в [67]. Его работа основана на генерации случайного четырёхзначного числа и выборке из четырёх SIMP-таблиц соответствующих частей предложения.

Этот способ не подходит для генерации текстов для поисковых систем, потому что текст состоит из заранее заготовленных словосочетаний и очень быстро перестает быть уникальным, но может служить вспомогательным при использовании других методов. Разработка таблиц является трудоемкой, созданная таблица подходит только для одной предметной области, что также накладывает ограничение на использование данного метода. Пример SIMP-таблиц приведен в виде табл. 1.1.

В столбце *A* приведены возможные начала предложений, представляющие собой вводные слова и обороты. В столбце *B* – подлежащие предложений; для снижения количества вероятных грамматических ошибок они предложены в единственном числе. Столбец *C* состоит из некоторых действий, которые совершают подлежащие, а столбец *D* – из универсальных окончаний предложения.

Таблица 1.1 – Пример SIMP-таблиц

№ п/п	A	B	C	D
1	Аналогично	отношение между двумя показателями нагрузки	свидетельствует о необходимости более тщательного анализа	интеграции и специализации.
2	Таким образом	формальное представление описываемого процесса	чрезвычайно усложняется, если не принять во внимание условие	сопровождения и поддержки.
3	Нетрудно видеть, что	необходимость проверки агрегата на корректность измерений	подразумевает более основательное использование теории	более тонкой аппаратной реализации.
4	Например,	постоянный поток эффективной информации	открывает весьма интересные перспективы	функционирования.
5	Итак,	тестирование устройства	признаёт значимость других систем и необходимость	предварительного отбора данных по определённым критериям.
6	Что касается нашей конкретной задачи, то	итог проведенных испытаний	указывает на пределы применимости	более детального исследования.

Если выбрать несколько случайных четверки чисел, например «5, 6, 4, 1», «3, 5, 1, 2», «6, 4, 3, 5», можно получить следующий текст:

*«Итак, итог проведенных испытаний открывает весьма интересные перспективы интеграции и специализации. Нетрудно видеть, что тестирование устройства свидетельствует о необходимости более тщательного анализа сопровождения и поддержки. Что касается нашей конкретной задачи, то постоянный поток эффективной информации*

*подразумевает более основательное использование теории предварительного отбора данных по определённым критериям».*

### **1.2.3 Метод генерации текста с использованием словарей**

Метод генерации с использованием словарей является наиболее трудоемким, но точным в плане соблюдения всех норм языка. Он требует специально подготовленных словарей с подробным перечислением характеристик слов и их форм, а также изучения порядка слов в предложении.

Метод с использованием словарей использует шаблоны известных ему форм языка – грамматических форм предложений и наличия зависимостей между словами. Любое предложение может быть представлено как формальное описание его компонент (лексем). Например, предложение «Мама мыла раму» представляет собой следующую форму: существительное женского рода единственного числа в именительном падеже, одушевленное + глагол прошедшего времени + существительное в винительном падеже. При заполнении данной формы случайными словами из словаря полученный результат может быть разнообразным: как «кошка ловила мышку», так и «мышка варила кошку». Таким образом, корректность создаваемых предложений напрямую зависит от подробности описания каждого компонента предложения, включая, например, тематику. Фактически генерация на основе словарей предполагает наличие некоторой формализованной грамматики, по которой строятся предложения.

Пример формальной грамматики  $G$  [68], описывающей формирование предложений, приведен ниже.

$$G = (N, T, P, S);$$

Начальный символ  $S = \langle \text{предложение} \rangle$ ;

Множество терминальных символов  $T = \{\text{невоспитанный, голодный, кабан, человек, жует, сопит, шумно, дико}\};$

Множество нетерминальных символов  $N = \{\langle \text{предложение} \rangle, \langle \text{словосочетание существительного} \rangle, \langle \text{словосочетание глагола} \rangle, \langle \text{прилагательное} \rangle, \langle \text{существительное} \rangle, \langle \text{глагол} \rangle, \langle \text{наречие} \rangle\};$

Множество правил вывода  $P$ :

1.  $\langle \text{предложение} \rangle \rightarrow \langle \text{словосочетание существительного} \rangle$   
 $\langle \text{словосочетание глагола} \rangle$
2.  $\langle \text{словосочетание существительного} \rangle \rightarrow \langle \text{прилагательное} \rangle$   
 $\langle \text{существительное} \rangle$
3.  $\langle \text{словосочетание существительного} \rangle \rightarrow \langle \text{существительное} \rangle$
4.  $\langle \text{словосочетание глагола} \rangle \rightarrow \langle \text{глагол} \rangle \langle \text{наречие} \rangle$
5.  $\langle \text{словосочетание глагола} \rangle \rightarrow \langle \text{глагол} \rangle$
6.  $\langle \text{прилагательное} \rangle \rightarrow \text{невоспитанный}$
7.  $\langle \text{прилагательное} \rangle \rightarrow \text{голодный}$
8.  $\langle \text{существительное} \rangle \rightarrow \text{кабан}$
9.  $\langle \text{существительное} \rangle \rightarrow \text{человек}$
10.  $\langle \text{глагол} \rangle \rightarrow \text{сопит}$
11.  $\langle \text{глагол} \rangle \rightarrow \text{жует}$
12.  $\langle \text{наречие} \rangle \rightarrow \text{шумно}$
13.  $\langle \text{наречие} \rangle \rightarrow \text{дико}$

Используя описанную грамматику, можно составить следующие предложение: «*Голодный кабан жует шумно*»; «*Человек сопит*»; «*Невоспитанный человек сопит дико*» и т.д.

Более детальное описание правил вывода, а также наполненный словарь терминальных символов позволяют создавать уникальные и корректные с точки зрения языковой грамматики тексты. В то же время проработка правил вывода и формирование словаря – достаточно долгий и

кропотливый процесс, требующий персонализации для отдельных предметных областей и разных жанров. Это причина того, что данный метод крайне редко используется при массовом порождении неестественных текстов.

#### 1.2.4 Метод генерации текста, основанный на синонимизации

Данный метод является одним из самых распространенных при создании контента для разнообразных ресурсов [69]. Синонимизация является эффективным инструментом как для массового порождения веб-контента, так и в целом для рерайта [70] или копирайтинга [71] – создания текста, к которому выдвигается требование уникальности, например эссе, реферата, отчета и т.д. Популярность метода обусловлена простотой применяемого алгоритма и используемых словарей.

Синонимизация подразумевает изменение текста путем замены отдельных лексем на схожие по смыслу (синонимы). Для реализации метода чаще всего используются программы-синонимайзеры, которые позволяют автоматизировать данный процесс, а также соответствующие словари синонимов [72]. Очевидно, что качество генерации, выражающееся в уникальности и корректности создаваемых текстов, напрямую зависит от качества и объема словаря.

Пример экземпляра словарной записи в формате [лексема | синоним 1, синоним 2, синоним 3, ...] выглядит следующим образом:

*Аккуратный | исправный, исполнительный, щепетильный, точный, тщательный, корректный, пунктуальный, скрупулезный, педантичный, неукоснительный, заботливый.*

Несмотря на то, что в используемых словарях используются отдельные записи для различных падежных окончаний и вариаций слова, не всегда удается учесть всю многогранность словообразования и



словоупотребления. В [73] приведен пример текстов с грамматическими и смысловыми ошибками, которые были созданы с помощью синонимайзера:

Исходный текст:

*«– Что ты видишь?*

*– Лошадь».*

Текст-результат:

*«– Что ты видишь?*

*– Конь».*

В следующем примере после работы синонимайзера утрачен смысл:

Исходный текст:

*«Спутниковая тарелка»*

Текст-результат:

*«Спутниковая миска»*

Для минимизации смысловых ошибок на уровне словосочетаний в современные синонимайзеры включен анализ  $n$ -грамм слов, который заключается в том, что заменяются только те слова, которые не нарушают общепринятые словосочетания, имеющие отдельный смысл.

Как было сказано ранее, качество генерации текста с помощью синонимизации напрямую зависит от наполненности словаря синонимов. Ниже приведены примеры естественного текста и созданных на его основе искусственных текстов. Искусственные тексты создавались с помощью синонимизации с использованием словарей разного объема.

Пример оригинального текста (фрагмент):

*«Полагаю, через десять лет экономика России будет такой же, как сейчас. Никаких радикальных изменений в её структуре мы, к сожалению, не увидим.*

*Это плохо, потому что мировая конъюнктура цен на нефть, газ и другие сырьевые ресурсы меняется, и не в нашу пользу. Это приведет к*

*тому, что Россия к 2026 году резко опустится в ранге стран с развитой экономикой.*

*Некоторые улучшения возможны разве что в сельском хозяйстве, IT, отдельных узких нишах машиностроения. Но на общем фоне деградации это не создаст нового качества экономики.*

*За счёт развития сельского хозяйства имеют шансы укрепиться Краснодарский край, Ставрополье, однако это также не скажется на общем состоянии экономики страны».*

Пример текста, который создан автоматически на основе исходного авторского фрагмента с помощью синонимизатора, использующего словарь синонимов для 180 тыс. слов:

*«Полагаю, сквозь десять лет экономика России довольно такой же, словно теперь. Никаких радикальных изменений в её структуре мы, к сокрушению, не увидим.*

*Это низко, потому который сделка конъюнктура валют для нефть, газ и другие сырьевые ресурсы меняется, и не в нашу выгоду. Это приведет к тому, который Россия к 2026 году явно опустится в ранге территорий с образованный экономикой.*

*Некоторые исправления допустимы неужели который в сельском хозяйстве, IT, частичных тесных нишах машиностроения. Однако для всеобщем фоне деградации это не создаст свежего качества экономики.*

*За счёт развития сельского хозяйства заключают шансы овладевать Краснодарский кромка, Ставрополье, впрочем это также не скажется для всеобщем положении экономики территории».*

Схожесть данного текста с оригиналом, рассчитанная с помощью алгоритма шинглов, составляет 49% для шингла из 1 слова, 24% – для шингла из 2 слов, 13% – для шингла из 3 слов. Так как уникальность данного текста составляет 51% (по лексическому соответствию), страница с ним с

высокой долей вероятности будет заблокирована поисковой системой для показа пользователю.

Пример текста, который создан автоматически на основе исходного авторского фрагмента с помощью синонимизатора, использующего словарь синонимов для 700 тыс. слов:

*«Считаю, посредством 10 года макроэкономика Российской федерации станет такого рода ведь, равно как в настоящее время. Практически никаких конструктивных перемен в ее текстуре я, к огорчению, никак не узнаем.*

*Это слабо, вследствие того то что всемирная ситуация стоимости в черное золото, голубое топливо и прочие сырьевые средства изменяется, и никак не в нашу с тобой выгоду. Данное повергнет к этому, то что Российская федерация к 2026 г. сильно снизится в ранге государств с сформированной экономикой.*

*Некоторые усовершенствования вероятны неужели то что в аграрном хозяйстве, IT, единичных ограниченных нишах машиностроения. Однако в совокупном фоне деградации данное никак не сформирует новейшего особенности экономики.*

*За счёт формирования аграрного хозяйства обладают возможности закрепиться Краснодарский область, Ставрополье, но данное кроме того никак не отразится в совокупном пребывании экономики государства».*

Схожесть данного текста с оригиналом, рассчитанная с помощью алгоритма шинглов, составляет 17% для шингла из 1 слова, 3% – для шингла из 2 слов, 0% – для шингла из 3 слов. Уникальность данного текста составляет достигает 83%, а повторов словосочетаний практически нет, страница с этим текстом будет предложена пользователю поисковой системой.

### 1.3 Методы определения массово порожденных текстов

Существующие решения в области определения текстов, сгенерированных автоматически, касаются в первую очередь выявления поискового спама. Поисковым спамом называются приемы искусственного поднятия рейтинга и продвижения интернет-ресурсов, то есть попытки обмана поисковых систем с целью повышения оценки релевантности страниц и, соответственно, изменения упорядоченности результатов поиска. Генерация и распространение поискового спама в сети различными способами монетизированы, и его доля от всего контента по разным оценкам составляет от 10–20 до 70% (в отдельных доменных зонах) [70, 75].

К инструментам создания поискового спама относят:

- дублирование веб-страниц;
- создание дорвеев;
- распространение ссылочной массы;
- манипуляции с текстом сайта, которые, в частности, связаны с использованием алгоритмов массового порождения текстов.

Паразитное воздействие поискового спама направлено в первую очередь на поисковые системы, с чем связаны и механизмы, используемые для его генерации и распространения. Следствием этого является то, что в его обнаружении заинтересованы в первую очередь компании, развивающие и поддерживающие поисковые системы. Так, крупнейшие из них имеют собственные научно-исследовательские лаборатории, занимающиеся решением задач обнаружения паразитного контента [76–80].

В работах Павлова А.С. и Райгородского А.М. с соавторами [81–83] предложен алгоритм определения текстового спама на основе оценки разнообразия тематик документа. Полученные Павловым А.С. с соавторами результаты по эффективности превосходят все известные алгоритмы поиска спама, кроме того, его алгоритм дает меньшую долю ошибок, как 1-го, так и

2-го рода, а также позволяет обрабатывать тексты в потоковом режиме. Методика основана на оценке характеристик текста, отличающих поисковый спам, – обилие ключевых слов, наличие скрытого текста, мелкого нечитаемого шрифта, злоупотребление тегами заголовков и т.д. Использование данных характеристик текста не позволяет использовать его методику для определения следов автоматического генератора на текстах, распространяемых как информационный контент, неблагоприятное воздействие которого направлено не на поисковые системы, а непосредственно на пользователей веб-ресурсов.

Также исследования веб-спама опубликованы зарубежными авторами C. Castillo, D. Donato et al. [16]. В основе предложенной ими методики – анализ веб-контента на наличие ссылок на определенные страницы и их распределения между страницами. В своей работе они приводят схемы взаимодействия ссылок для паразитных «спам»-страниц и аналогичные – для «нормальных» страниц. Данная методика обнаружения автоматически сгенерированного текста также имеет в своей основе особенности веб-спама и неприменима для иных текстов.

В работе Зайцева А.А. с соавторами [84] рассмотрен метод оценки качества текста на основе реферирования. Метод предполагает, что тексты, созданные с применением синонимизаторов или средств автоматического перевода, обладают меньшей тематической устойчивостью, поэтому при повышении компрессии «сжатия» реферата у них с большей скоростью уменьшается его размерность. Авторами выведен ряд эвристических правил, которые позволяют исключить тексты низкого качества, тем самым доказывая, что тематические признаки текста позволяют выявить связные, информативные тексты.

Определение текстов, представляющих собой машинный перевод с одного естественного языка на другой, – еще одна задача, связанная с

определением происхождения текста. Такие тексты могут быть использованы и в качестве материала для создания поискового спама, и для распространения в качестве информационного контента. Для выявления таких текстов существует ряд специальных алгоритмов, например BLEU, METEOR, Perplexity, оценивающих качество машинного перевода через соответствие характеристик текста на лексическом уровне некоторым эталонным показателям естественного языка. Данные алгоритмы опробованы в ряде работ для решения различных прикладных задач. Например, в работе R. Aharoni с соавторами [85] предлагается расширить алгоритм BLEU рассмотрением текстов также на уровне предложения. Авторами показано, что для некоторых языков данный подход дает более точные показатели метрики BLEU.

В исследованиях текстов, созданных автоматически, часто уделяют внимание тематическим и семантическим свойствам текста. Для оценки текстов в таком случае часто используются методы автоматической рубрикации. Например, в работе авторов Е.В. Дунаева и А.А. Шелестова [22] приведены подходы к рубрикации текстовых данных на основе алгоритма PrTFID. Автоматическая рубрикация может быть использована для выделения значимых участков текста при определении автоматически сгенерированных текстов.

Однако кроме неоспоримой вычислительной эффективности – рубрикация текстов в автоматическом режиме занимает доли секунды – есть и определенные затруднения в использовании таких алгоритмов. При автоматической обработке текстовых потоков могут возникнуть проблемы анализа языкового материала, контекста употребления тех или иных слов, требующие привлечения обширных знаний о языке и предметной области, которые трудно описать в существующих программных системах автоматической рубрикации. Также для автоматической рубрикации

необходимо создание образа рубрики как некоторого формального выражения на основе слов или терминов реальных текстов. В масштабах обработки электронного контента это может быть затруднительно и потребует привлечения большого количества экспертов в различных областях знаний.

#### **1.4 Алгоритм определения искусственно созданных текстов**

Автоматическая обработка текста при атрибуции в целом и определении искусственных текстов в частности, как правило, состоит из нескольких этапов:

1. Считывание электронного контента (материала) из источника, которым в разных случаях может выступать веб-сайт, специальный сервис, файл некоторого формата и т.д.;

2. Выделение текстовой информации из считанного материала, при котором извлекается только исследуемые текстовые данные для последующей обработки;

3. Анализ текстовых данных, который включает в себя исследование различных характеристик, в том числе частотных, синтаксических, грамматических и др.;

4. Атрибуция текста, включающая расчеты, которые позволяют принять решение об условиях их создания текста, в частности – был ли текст создан человеком или создан автоматически с помощью специального алгоритма генерации;

5. Вывод принятого решения, является ли текст естественным или искусственным.

Алгоритм определения искусственно созданных текстов, представленный в виде блок-схемы представлен на рис 1.1.

В диссертационной работе автором рассматривается только обработка текстовых данных, включающая анализ текста и его атрибуцию, которые в совокупности позволяют определить искусственно созданные тексты.

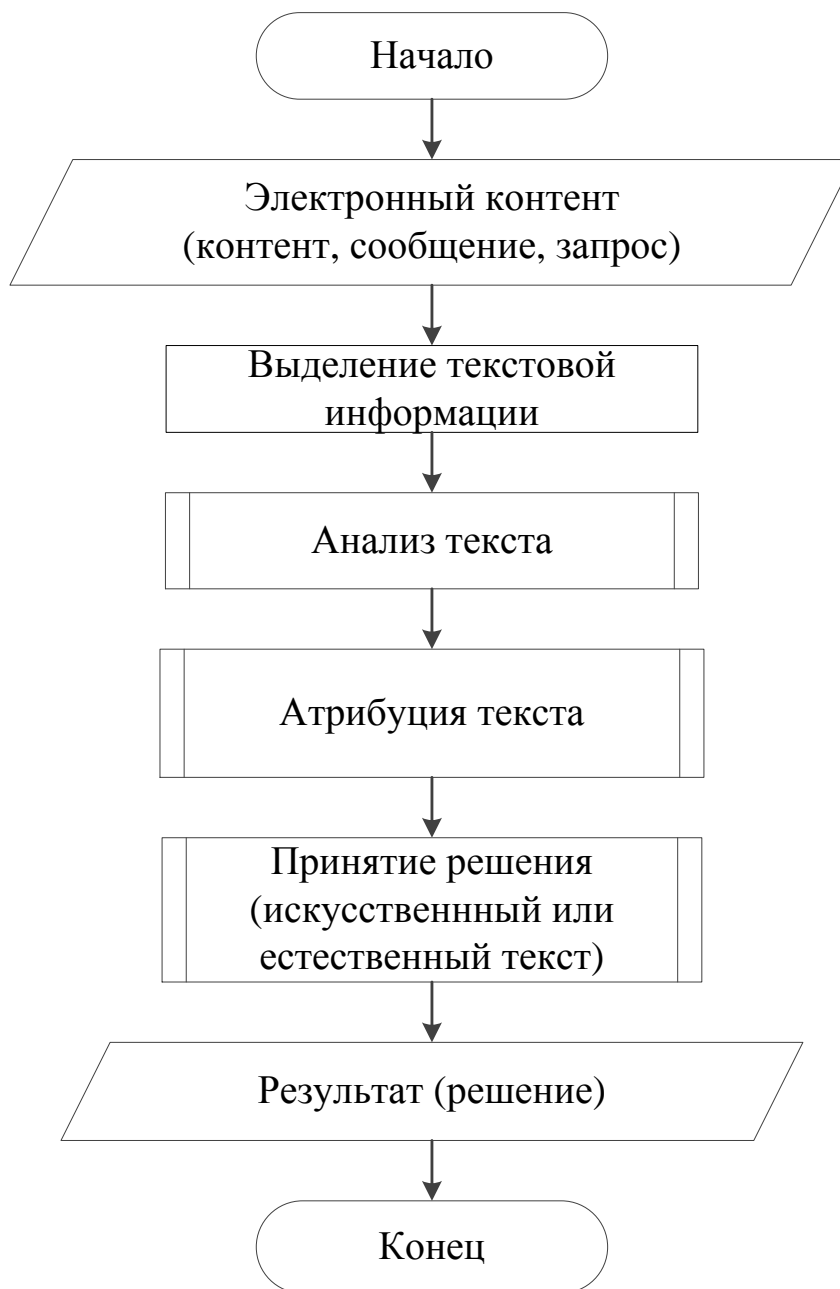


Рисунок 1.1 – Блок-схема алгоритма определения искусственно созданных текстов



## 1.5 Выводы

Задача определения искусственно созданных текстов связана с необходимостью выявления характерных языковых особенностей в рассматриваемых текстах. Поскольку определение происхождения текста востребовано в различных сферах деятельности, а экспертный анализ авторского стиля является трудоемким и длительным процессом, актуальными являются создание и развитие формальных методов решения данной задачи.

На сегодняшний день среди множества работ ученых всего мира не представлено общепризнанного и эффективного решения задачи определения искусственных текстов. Кроме того, рассматриваемая в данном исследовании задача однозначно не классифицирована: она имеет схожие черты с задачей верификации авторства, а также с задачей определения авторства [86]. По этой причине была предпринята попытка проанализировать применение существующих методов текстовой атрибуции для решения задачи определения искусственно созданных текстов.

По итогам проведенного анализа существующих методов и алгоритмов определения авторства и текстов, созданных искусственно, то есть с применением программных алгоритмов, были сделаны следующие выводы:

1. Существующие методы определения неестественных текстов позволяют решать задачи выявления поискового спама, а также текстов, являющихся машинным переводом текста с одного естественного языка на другой.

2. Из-за наложенных ограничений и используемых особенностей рассмотренных текстов существующие методы не могут быть использованы в задаче определения текстов, созданных с помощью синонимизаторов или других средств автоматической генерации.

3. Для достижения большей точности в определении связности текста необходимо рассматривать все его лингвистические уровни, в том числе синтаксический, на уровне предложений.

4. Качество текста напрямую связано с его тематическими свойствами, то есть связанные с ними характеристики текста должны быть включены в инварианты.

5. При формировании набора текстовых характеристик, используемых для определения неестественных текстов, следует отталкиваться от глобальных свойств текстов.

Приведенный автором алгоритм определения искусственно созданных текстов показывает последовательность действий для принятия решения о способе создания исследуемого текстового материала. В ходе диссертационного исследования автором предполагается создание методического обеспечения, которое позволит определить искусственно созданные тексты, представленные в виде текстовых файлов.

Таким образом, необходимо разработать метод определения текстов, сгенерированных автоматически, представляющих собой информационный веб-контент, а также алгоритм анализа характеристик данных текстов. Для исследования предлагается использовать методы статистического анализа, так как они, в отличие от методов машинного обучения, позволяют явно выделить устойчивость значения характеристик текста и применить к ним известные средства статистического анализа, например корреляционной оценки, и в то же время являются эффективным методом текстовой атрибуции.

При решении данной задачи необходимо учесть достижения исследований по оценке качества текстовых произведений и определения поискового спама, а также в области выделения характерных черт естественных текстов [19] и автоматической атрибуции текстов [46, 87–90], которые косвенно относятся к решению задачи определения массово порожденных текстов.

## **2 РАЗРАБОТАННЫЕ ИНВАРИАНТЫ ИСКУССТВЕННО СОЗДАНЫХ ТЕКСТОВ**

Согласно алгоритму определения искусственно созданных текстов, представленного автором в п. 1.4, после выделения текстовых данных из исследуемого материала следует их анализ. Под анализом текста в рамках решаемой задачи понимается вычисление значений выделенных характеристик текста, которые выделены в инвариант. Инвариантом в данном случае является набор характеристик текстов и их значений, позволяющий отличить естественные и искусственные тексты.

Формирование инвариантов классов текстов является важной задачей атрибуции, так как инварианты используются в дальнейшем для принятия соответствующего решения о тексте. В то же время данная процедура носит субъективный характер в связи с тем, что выполняется исследователем-экспертом. Автором предлагается формализованный подход к формированию инвариантов, представленный в виде алгоритма и основанный на исследовании лингвистических особенностей исследуемых классов текстов. На основе авторского алгоритма предлагается создать инварианты классов естественных и искусственных текстов, которые впоследствии будут использованы для автоматического определения искусственно созданных текстов. Результаты исследования характеристик указанных классов текстов и создания их инвариантов на основе предлагаемого алгоритма опубликованы автором в [91].

### **2.1 Классификация текстовых произведений на основе инвариантов**

По примеру методов, применяемых при решении задач идентификации авторства текстов, для определения их происхождения

необходимо опираться на ряд текстовых характеристик и анализировать их значения. В исследованиях, посвященных автоматизированному определению авторства, была доказана возможность формирования собственного инварианта для отдельного автора или группы авторов. Однако для идентификации естественных и искусственных текстов необходимо исследовать значения текстовых характеристик этих двух классов текстов на возможность их различительной способности. В рамках решаемой задачи можно выделить 2 класса текстов: естественные и искусственные. Под первыми понимаются тексты, созданные человеком, под вторыми – созданные автоматически с помощью специального программного алгоритма. При этом класс искусственных текстов также может быть разбит на несколько других – по признаку применения того или иного метода при генерации. Для отнесения входного текста к одному из исследуемых классов необходимо сформировать соответствующие инварианты.

Классификация текста, то есть определение его соответствия некоторому способу создания, автору (группе авторов) или иному классу, может быть представлена в формальном виде, основанном на теоретико-множественном подходе.

$T$  – множество всех текстов.

$K$  – множество классов текстов,  $|K| = n$ .

$A = \{a_1, a_2, \dots, a_n\}$  – множество инвариантов классов текстов в рамках решаемой задачи,  $|A| = n$ .

$T_a$  – множество текстов, которым сопоставлен некоторый инвариант, то есть тексты известного происхождения или авторства.

$T'$  – множество текстов, которым не сопоставлен инвариант то есть совокупность текстов неизвестного класса.

При этом выполняется:  $T = T_a \cup T'$ ;  $T_a \cap T' = \emptyset$ .

$X$  – конечное множество исследуемых текстовых характеристик, множество определяется следующим образом:  $X = \{x \mid x \text{ – изучаемая текстовая характеристика}\}; |X| = m$ .

Инвариант  $a_j \in A$  представляет собой массив упорядоченных пар вида:  $\langle \text{текстовая характеристика } x_i \in X; \text{ значение текстовой характеристики } x_i \text{ для данного инварианта } z_{ij} \rangle$ . То есть  $a_j = (\langle x_1, z_{1j} \rangle, \dots, \langle x_i, z_{ij} \rangle, \dots, \langle x_m, z_{mj} \rangle)$ , где  $i = 1, \dots, m; j = 1, \dots, n; z_{ij}$  – некоторое числовое значение, формат и диапазон которого выбраны в соответствии с текстовой характеристикой, в отдельных случаях в качестве  $z_{ij}$  могут выступать диапазоны значения характеристики:  $z_{ij} = [z_{ij \min}; z_{ij \max}]$ .

Тогда на декартовом произведении множества текстов  $T$  и множества инвариантов  $A$  может быть задано бинарное отношение  $R \subset T \times A$  такое, что выполняется  $tRa$ , если некоторый текст  $t \in T$  соответствует инварианту  $a \in A$ , то есть текст  $t$  относится к классу, которому соответствует инвариант  $a$ , или текст  $t$  написан автором, которому соответствует инвариант  $a$ .

Учитывая все предшествующие обозначения, можно записать условие того, что некоторому входному тексту  $t \in T$  соответствует инвариант  $a \in A$ . Отношение  $tRa$  выполняется, если значения текстовых характеристик исследуемого текста  $t$  соответствуют или приближены в определенной степени к значениям характеристик  $x_i \in X$  инварианта  $a$ . При этом степень приближенности значений устанавливается автором метода в каждом конкретном случае и должна быть обоснована экспериментальными данными.

Очевидно, что инвариант автора и инвариант метода создания текста имеют существенные отличия. Главным отличием инварианта метода создания текста (генератора искусственных текстов) от инварианта автора или группы авторов является то, что первый должен быть применим для искусственного текста, созданного на основе любого авторского материала,

который был использован генератором в качестве исходного текста (в случае для генераторов, создающих текст на основе существующего исходного материала). Учитывая данный факт, необходимо в первую очередь выделить набор характеристик, значения которых позволили бы отличить тексты, написанные человеком, и тексты, созданные с помощью программных генераторов. Характеристики должны обладать свойствами, достаточными для их использования при автоматизированном определении происхождения текстов. Под такими свойствами в теории атрибуции понимаются массовость, устойчивость и различающая способность текстовых характеристик.

## **2.2 Предложенный алгоритм формирования инвариантов классов текстов**

В классической задаче атрибуции – установлении авторства – инвариант, на основе которого идентифицируется автор, представляет собой набор значений характеристик текста определенного лица [92]. Для создания такого набора существует несколько подходов. В случае с идентификацией искусственных текстовых произведений инвариантом является набор значений характеристик текста, с помощью которых может быть установлена причастность данного генератора к происхождению входного текста [93, 94].

Многими учеными предпринимались попытки смоделировать подход к формированию набора характеристик текста, составляющих инвариант. В ранних работах, посвященных обработке и классификации текстовых произведений, в основе выбора характеристик лежал либо интуитивный подход, либо случайный перебор. Более поздние исследования использовали накопленные знания о различительной способности тех или иных характеристик, развивая и совершенствуя их. Однако особенность

формирования инвариантов состоит в том, что наборы исследуемых характеристик зависят в первую очередь от непосредственно решаемой задачи, и процесс формирования набора должен отталкиваться от задачи классификации.

На сегодняшний день существуют некоторые стандартные наборы характеристик, элементы которых чаще всего используются для расчета инварианта при решении тех или иных задач атрибуции. На использовании таких наборов основываются модели создания инвариантов текстов при решении задач классификации.

В работе [88], посвященной определению авторства текстов, приводится методика, в которую включено описание процесса создания инвариантов в виде блока функциональной модели (приведена на рис. 2.1).

Процесс формирования авторского стиля в данном случае описан следующим образом:

– на вход подаются доступные признаки текста, которые пользователь объединяет в некоторую группу признаков текстов;

– данная группа признаков текста, а также множество текстов известного авторства используются для формирования модели авторского стиля, то есть инварианта с учетом требований к точности определения автора.

Формированию инвариантов согласно предложенной Романовым А.С. методики основана на использовании известных наборов характеристик вне зависимости от особенностей решаемой прикладной задачи. При таком подходе значительно возрастает вычислительная сложность расчетов, так как количество всевозможных характеристик может составлять несколько тысяч. Кроме того возникает риск упущения каких-либо характеристик текста, которые отсутствуют в стандартных наборах, но в конкретном случае могут

обладать различительной способностью, что может стать причиной увеличения количества ошибок при атрибуции текстов.

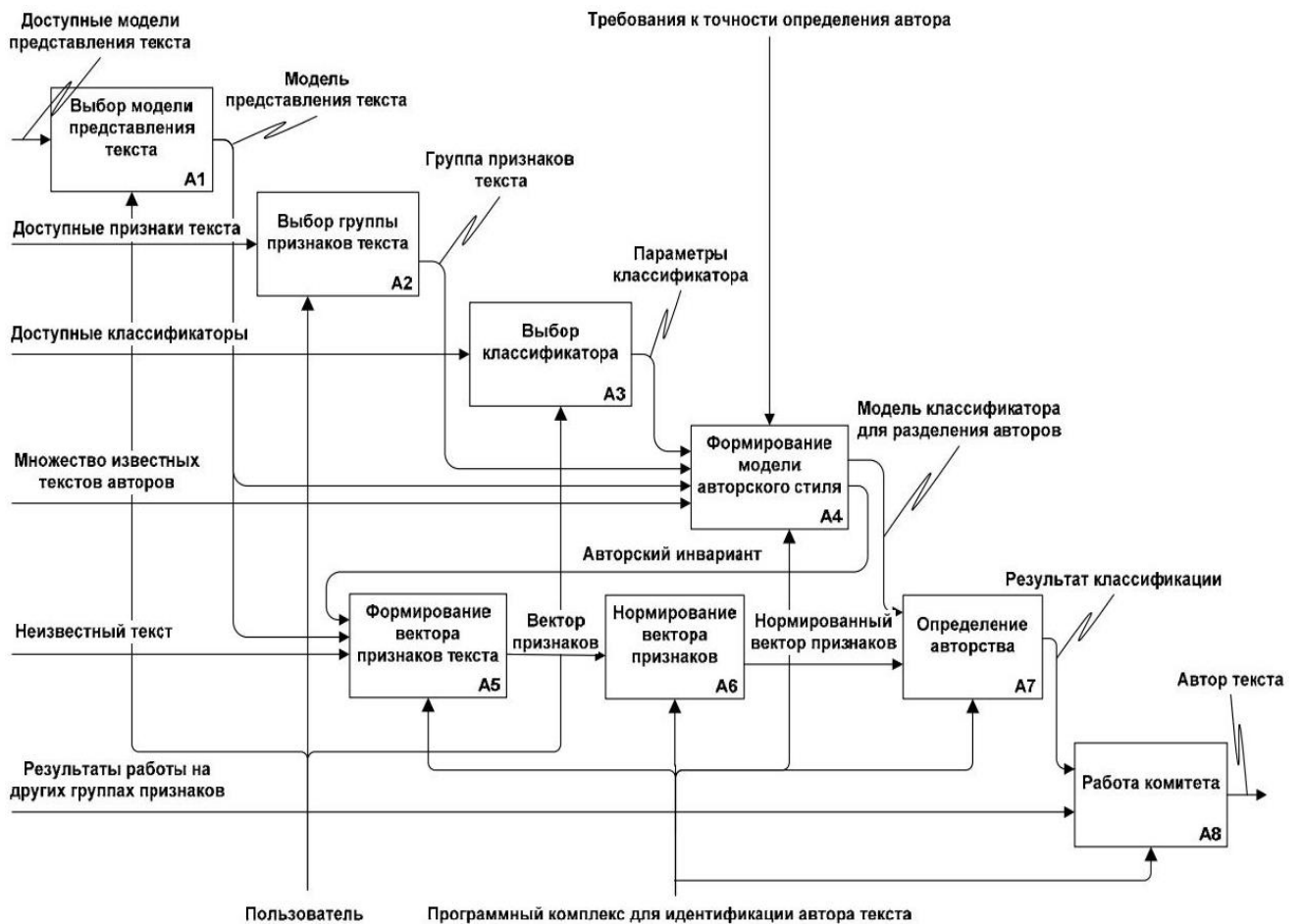


Рисунок 2.1 – Методика Романова А.С. для идентификации автора неизвестного текста

Особенность формирования инвариантов в задачах атрибуции текстов должна состоять в том, что наборы исследуемых характеристик зависят, прежде всего, от непосредственно решаемой задачи, а процесс формирования набора отталкивается от задачи классификации. Также следует отметить, что любая задача, связанная с атрибуцией текста, является междисциплинарной и связывает знания в области лингвистики, стилометрии, информатики, статистики и, возможно, других областей знаний. Таким образом,



формирование инвариантов классов текстов должно быть основано на лингвистических особенностях языка и текста.

В соответствии с приведенными особенностями исследования текстов автором был предложен алгоритм формирования инвариантов классов текстов, который отличается наличием процедуры выделения качественных признаков, различающих исследуемые классы текстов. Таким образом, при формировании набора исследователя основывается на лингвистических особенностях рассматриваемых классов текстов. Алгоритм является универсальным и может быть применен при решении любой задачи, связанной с классификацией текстов.

Предложенный алгоритм позволяет сформировать множество количественных характеристик текста на основе выделенных качественных признаков, которые выделяются на основе лингвистических особенностей исследуемых классов текстов.

Блок-схема алгоритма, разработанного автором, представлена на рис. 2.2. Входными данными для формирования инвариантов согласно нему являются:

– множество классов текстов  $K$ , для которых необходимо сформировать инварианты; классы разделяют множество текстов  $T$  на соответствующие подмножества по некоторому устойчивому и однозначному свойству, например по способу создания, автору, некоторому признаку автора;

– наборы текстов  $T_i$  ( $i = 1..n$ ) для каждого класса из множества  $K$ ; объем текстового материала для одного набора, согласно рекомендациям, принятым в теории текстовой атрибуции [23], не должен быть менее 16 тыс. символов.

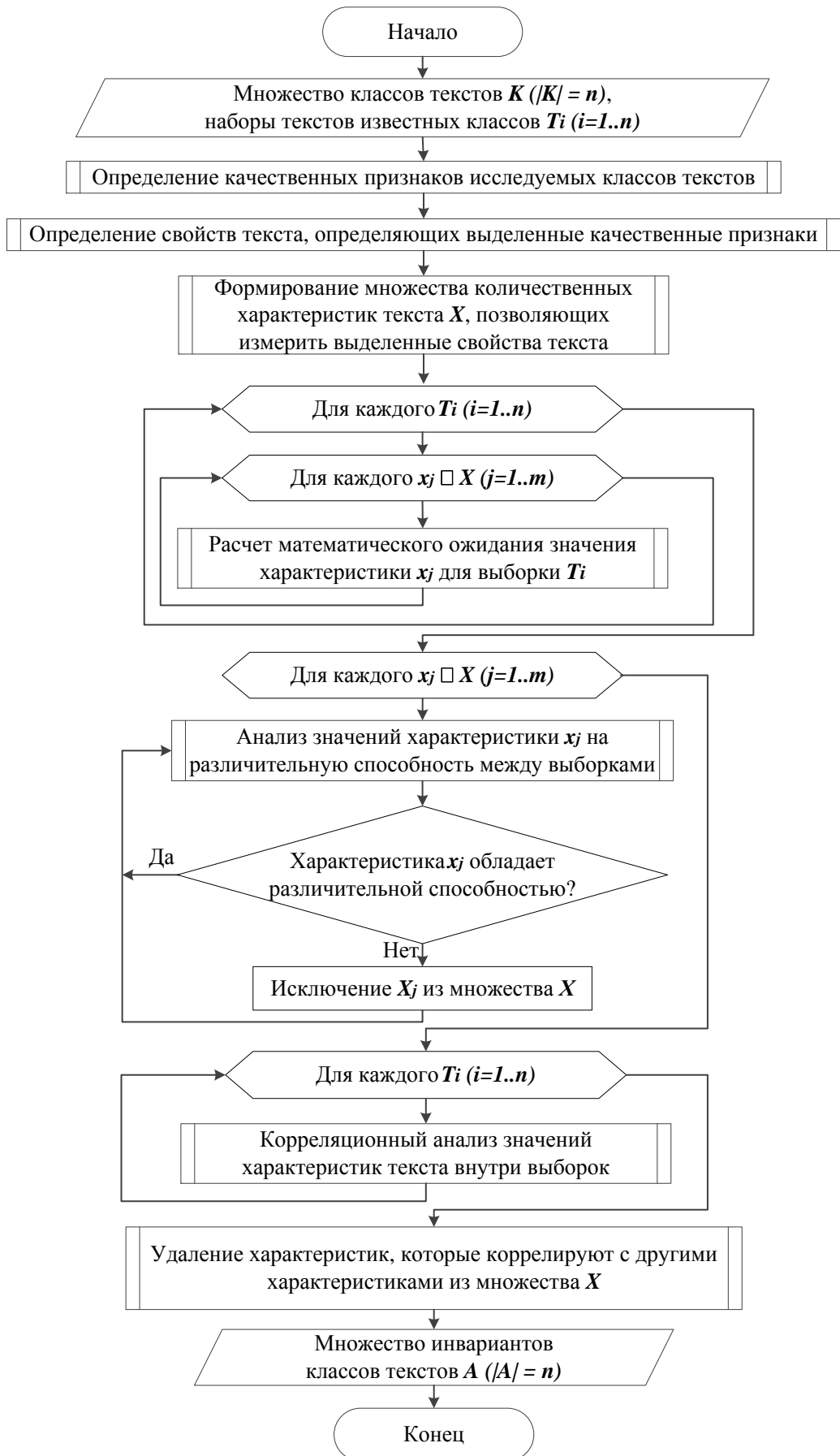


Рисунок 2.2 – Предложенный автором алгоритм формирования инвариантов

В ходе формирования инвариантов согласно предложенному алгоритму происходит последовательное выполнение следующих процедур:

1. Определение качественных признаков исследуемых классов текстов с учетом лингвистических особенностей языка. Данные признаки должны отражать основные отличия между текстами разных выборок. Примерами данных признаков могут выступать:

- эмоциональная окраска;
- стилевое единство;
- соответствие жанру, стилю;
- соответствие текста времени его создания;
- логичность повествования;
- связность текста и т.п.

Под лингвистическими особенностями языка понимаются обоснованные учеными-филологами и учеными-лингвистами научные факты, суждения о классификациях, взаимосвязях и других проявлениях языка и всех его компонент.

2. Уточнение свойств текста, которые позволяют определить проявление выделенных качественных признаков. Данный этап также основан на использовании знаний в области лингвистики. Свойства обычно обобщают особенности текста на уровне одних и тех же текстообразующих компонент.

3. Формирование множества количественных характеристик  $X$ , позволяющих оценить проявление выделенных на предыдущем этапе свойств текста. Множество характеристик  $X$  полностью обусловлено лингвистическими особенностями исследуемого языка и классов, которые анализируются исследователем-экспертом.

Последующие этапы алгоритма носят расчетный характер. Расчеты производятся на основе соответствующего математического аппарата,

применимость которого должна быть обоснована опытом исследований в рассматриваемой области, а также логическими суждениями и внутренней непротиворечивостью производимых вычислений.

4. Расчет математического ожидания значения характеристик  $x_j \in X$  для текстов из набора  $T_i$  производится циклично для каждой характеристики, для каждой выборки ( $i = 1..n, j = 1..m$ ).

5. Анализ значений характеристик  $x_j \in X$  на их различительную способность между исследуемыми наборами текстов производится в цикле при  $j = 1..m$ . Для измерения различительной способности применяется мера, позволяющая оценить различие между значениями характеристики для наборов текстов разных классов.

6. В случае если характеристика  $x_j \in X$  не удовлетворяет условию различительной способности, она исключается из множества исследуемых характеристик.

7. Анализ характеристик на корреляционную зависимость проводится для всех  $x_j \in X$  внутри каждого набора текстов одного класса. Расчеты на данном этапе производятся на основе соответствующего математического аппарата, применимость которого должна быть обоснована опытом исследований в рассматриваемой области, а также логическими суждениями и внутренней непротиворечивостью производимых вычислений.

8. Удаление характеристик, коррелирующих с другими характеристиками внутри наборов текстов, позволяет окончательно сформировать множество характеристик  $X$ .

Результатом выполнения всех этапов процесса являются инварианты классов текстов, объединенные множеством  $A$  и представляющие собой наборы численных значений характеристик текста множества  $X$ .

Процесс формирования инвариантов классов текстов, представленный в виде диаграммы в нотации IDEF0, изображен на рис. 2.3.

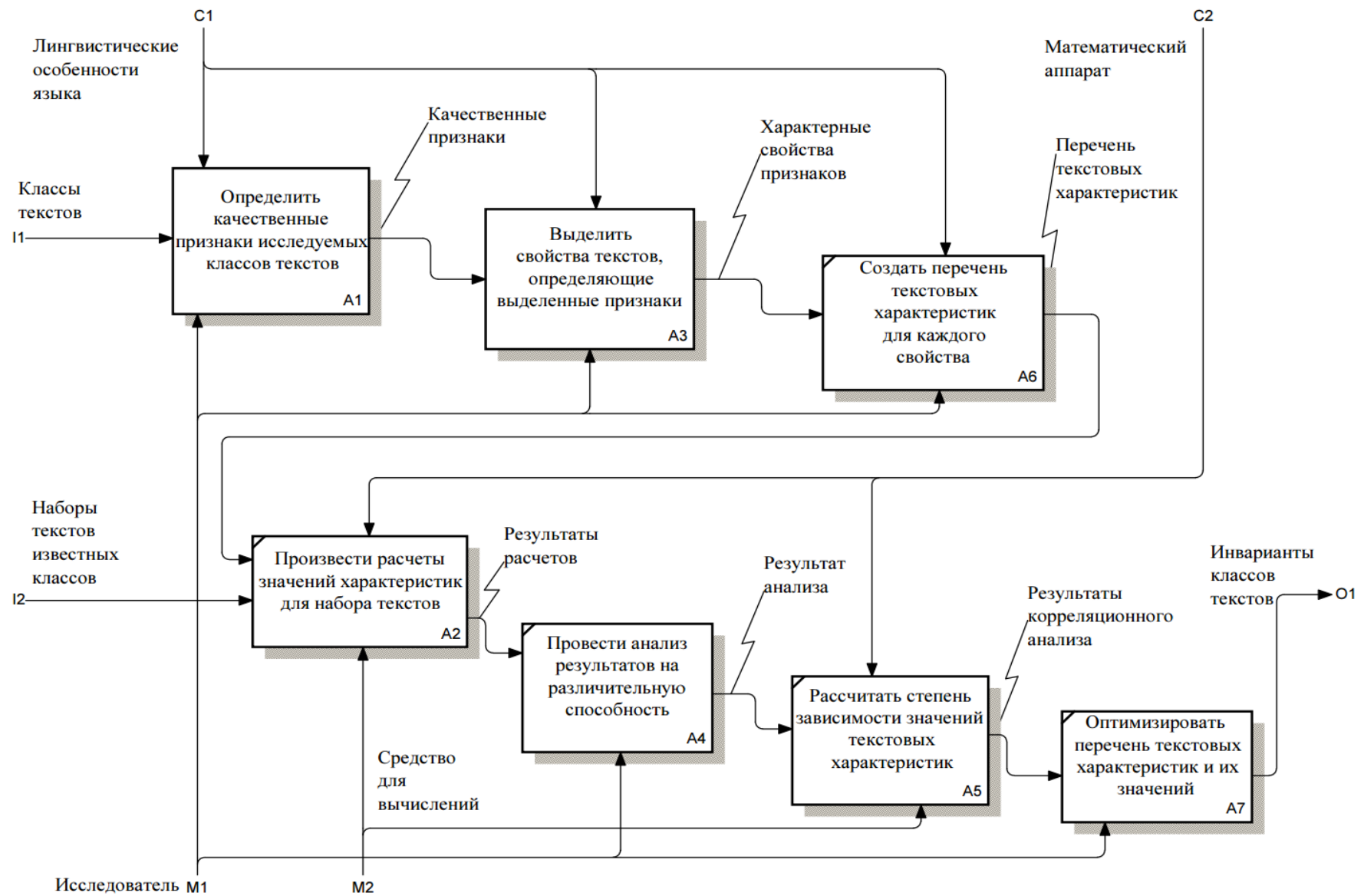


Рисунок 2.3 – Диаграмма процесса формирования инвариантов классов текстов

Разработанный и представленный в виде блок-схемы, описания и диаграммы алгоритм формирования инвариантов классов текстов отличается от существующих представлений использованием качественных и уточняющих их количественных текстовых характеристик, основанных на лингвистических особенностях языка.

### **2.3 Формирование инвариантов искусственных и естественных текстов на основе предложенного алгоритма**

Для формирования инвариантов необходимо в первую очередь выделить набор исследуемых классов и составить наборы текстов, относящиеся к данным классам. В настоящем исследовании были рассмотрены наиболее распространенные методы генерации, которые в то же время позволяют создать тексты с наилучшими показателями уникальности, – метод, основанный на цепи Маркова и синонимизация.

Применение данных методов генерации текста позволяет достичь уникальности свыше 80 %, в среднем тексты, созданные на основе них, имеют уникальность около 65 %. Такие высокие показатели достигаются за счет замены лексем и их последовательности в тексте, которое слабо выражено, например, в методе с использованием SIMP-таблиц. Достоинством методов синонимизации и Марковских цепей в рамках задачи создания искусственного Интернет-контента, электронных сообщений и т.п., является также простая программная реализация, не требующая значительных вычислительных ресурсов. В то время как методы на основе SIMP-таблиц и на основе словарей требуют разработки специальных шаблонов.

Были созданы три выборки по 3 210 текстов:

– в первую выборку вошли естественные тексты, созданные человеком (объем выборки составляет 10,7 млн символов);

– во второй – экземпляры, сгенерированные автоматически с помощью метода на основе цепей Маркова (объем выборки составляет 14,1 млн символов);

– в третий – сгенерированные на основе синонимизации с использованием словаря, содержащего синонимы к 700 тыс. слов (объем выборки составляет 12,3 млн символов).

Естественные тексты представляют собой публицистические статьи информационного характера длиной от 1 000 до 5 700 символов. Экземпляры искусственных текстов генерировались на основе текстов из первой выборки. В работах [23, 89] отмечается, что указанные объемы выборок считаются достаточными для обучения в соответствии с опытом формирования инвариантов при анализе текстов различных авторов. Показатели уникальности сгенерированных текстов по сравнению с оригинальными экземплярами в рамках представленных выборок приведены в табл. 2.1. Уникальность была рассчитана с помощью алгоритмов шинглов [65] для шингла из 1 слова.

Таблица 2.1 – Уникальность сгенерированных текстов

Выборка искусственных текстов	Среднее значение уникальности сгенерированных текстов (расчет произведен с помощью алгоритма шинглов)
Тексты, созданные с помощью метода цепей Маркова (с использованием программного средства Delirium [95])	69,13 %
Тексты, созданные с помощью метода синонимизации (с использованием программного средства Article Clone Easy [96])	64,70%

Согласно предложенному в п. 2.2 алгоритму формирование инварианта для решения задачи определения происхождения текстов представляет собой последовательность нескольких этапов:

- 1) определение качественных признаков, по которым различимы тексты разных классов;
- 2) определение свойств текста, на которые влияют выделенные качественные признаки;
- 3) формирование множества количественных характеристик текста, которые позволяют измерить проявление свойства текста;
- 4) расчет значений математического ожидания характеристик текста;
- 5) проверка различительной способности характеристик;
- 6) исключение характеристик, не обладающих различительной способностью;
- 7) проверка корреляционной зависимости характеристик;
- 8) удаление характеристик, коррелирующих с другими характеристиками в наборе;
- 9) формирование инвариантов как наборов значений выделенных характеристик текста.

### **2.3.1 Определение качественных признаков выделенных классов текстов**

Естественный текст отличается от искусственного своей связностью в рамках межфразовых единств, а также цельностью, то есть глобальной связью компонентов текста на содержательном уровне [97].

Связность текста обнаруживается как связанность его компонентов, согласованность фраз, предложений и других текстообразующих средств. Текст является связным, когда четко фиксируются структурные показатели связи, в качестве которых в языке выступают такие средства, как союзы, вводные слова и сочетания, плавные переходы от темы к теме и т.п. Связь также может обнаруживаться без специальных словесных сигналов связи – характерным расположением слов и словосочетаний в тексте. Цельность в тексте поддерживается ключевыми словами и их заместителями, ее возможно



определить, последовательно исследовав текстовое произведение через понятийную последовательность в изложении [98].

Таким образом, связность и цельность являются неперенными признаками текста, которые проявляются в целесообразно построенном человеком тексте и отличают его от массово порожденных экземпляров.

### **2.3.2 Определение свойств текста, определяющих проявление качественных признаков**

В работах [97, 98], посвященных лингвистическим особенностям естественных текстов, выделяют следующие классы свойств текста, обеспечивающие его связность и цельность:

- символные (связанные с наличием букв, буквосочетаний, цифр, символов пунктуации, математических символов);
- лексические (связанные с наличием слов и словосочетаний, «словарный запас»);
- синтаксические (связанные с конструкциями предложений в тексте, локальными или глобальными синтаксическими особенностями текста);
- семантические (связанные с оценкой мер семантического сходства и связанности текстообразующих средств);
- тематические (связанные с соответствием используемых текстообразующих средств тематике текста).

В системах автоматизированного определения авторства или диагностики чаще всего используются символные и лексические свойства текстов. Это в первую очередь обусловлено вычислительной простотой расчетов, связанных с определением значений текстовых характеристик. В свою очередь синтаксические и семантические анализаторы, а также тематические рубрикаторы предъявляют дополнительные требования к

разметке текста и зависят от качества используемых справочников. Анализ таких свойств характеризуется высокой вычислительной сложностью, так как менее формализован и зависит от многообразия форм языка.

### **2.3.3 Формирование множества количественных характеристик**

Для выделения характеристик текста, которые определяют выделенные свойства, необходимо подробно рассмотреть особенности каждой из групп свойств и возможность их применения для определения искусственных текстов.

Символьные свойства текста активно используются в работах [4, 7, 46] для определения авторства, чаще всего это частоты букв и буквосочетаний, знаков пунктуации, эмодиконов. В случае с искусственно созданными текстами численные значения данных характеристик в большей степени зависят от исходного текста либо от базы используемых синонимов, поэтому они не могут отразить влияние генератора на текст. Однако замена слов синонимами и перестановка слов с помощью Марковской цепи могут отразиться на длинах слов, так как происходит замена простых, часто употребляемых слов на нераспространенные синонимы, которые чаще всего имеют большую длину. Поэтому следует исследовать влияние генератора на длины слов в тексте, количество коротких и длинных слов.

Лексические свойства текста являются очень важными в формировании инварианта генераторов искусственных текстов. Их преимуществом является вычислительная простота расчетов, а также очевидность в интерпретации результатов. Оба метода генерации, рассматриваемые в диссертационной работе, основаны на манипуляциях с лексемами: замене слов синонимами или изменении положения слов в тексте, основанном на вероятностном распределении переходов от слова к слову в исходном тексте.

Предполагается, что именно лексические свойства станут основой идентификации искусственных текстов. Массовая замена слов синонимами с использованием словаря приводит к тому, что в сгенерированном тексте встречаются малоупотребимые слова и словосочетания, а количество распространенных слов, напротив, падает. К наиболее часто встречающимся в русских текстах словам в первую очередь относятся служебные слова и местоимения как незаменимые связки в письменной речи. Влияние характеристик, связанных с количеством и распределением служебных слов, на авторский стиль текста, описано в работе [23]. В определении происхождения текста данные характеристики могут также оказаться информативными, так как изменения структуры текста, лексического набора напрямую влияет на наличие служебных слов, что показано в [99].

Синтаксический анализ предложений искусственного и естественного текстов свидетельствует о том, что синтаксические свойства языка и значения характеристик, связанных с ними, значительным образом различаются. Это связано с неграмотностью построения предложений и фраз генераторами, сложностью формализации правил грамматики языка в его многообразных проявлениях, наличием множества исключений, а также вычислительной и алгоритмической сложностью синтаксической проверки создаваемых текстов. Однако эти же обстоятельства затрудняют и расчет (измерение) текстовых характеристик, отражающих синтаксис, поэтому автором были выбраны характеристики, которые можно достоверно рассчитать, не прибегая к трудоемким алгоритмам синтаксического анализа, и которые в то же время отражают особенности естественной письменной речи русского языка. К ним были отнесены количество предложений в тексте, процент сложноподчиненных предложений [5].

Семантическая связь и единство в тексте также является слабоформализуемым свойством. Учитывая особенности алгоритмов

генерации текстов, можно заключить, что одной из нескольких причин потери семантической связи внутри текста является неточная, иногда неуместная, двусмысленная замена слов синонимами, что приводит к разобщению фраз, предложений [100]. Таким образом, в рамках решаемой задачи семантические и лексические свойства текста пересекаются, так как семантическая связь внутри текста теряется из-за замены или перемещения слов. Исследуя количество часто употребляемых слов и словосочетаний (*n*-грамм) в искусственных текстах относительно естественных, можно определить наличие количественной зависимости семантических свойств данных двух классов текстов. Для того чтобы набор часто употребляемых слов не зависел ни от жанра исследуемых текстов, ни от особенностей автора, генератора и тому подобного, были использованы данные Национального корпуса русского языка и рассчитанные для корпуса частоты словоформ и словосочетаний [101]. Всего корпус содержит 192 689 044 словоформы, он включает в себя тексты различных временных периодов с середины VIII века до настоящего времени, но основу составляют современные текстовые произведения различных жанров, то есть выборку текстов для вычисления частоты употребления слов можно считать репрезентативной, а данные корпуса использовать в диссертационной работе.

Тематические свойства текста также подвергаются влиянию при воздействии алгоритмов генераторов. Причины в данном случае сходны с описанными выше. Для оценки элементов некоторого текста на сходство тематики могут быть использованы алгоритмы автоматической рубрикации текстов, тематические классификаторы и т.п. Следует принимать во внимание, что большая часть исследований по автоматической рубрикации и классификации проводится на текстах объемом свыше 5 000 символов, это диктуется в том числе областью применения подобных алгоритмов. Для оценки тематического единства текста в решении задачи настоящей

диссертационной работы предлагается использовать количественную характеристику, описывающую сходство тематик в 3 разных частях текста.

Учитывая это, был сформирован перечень количественных характеристик, позволяющих измерить проявление выделенных свойств текста:

- средняя длина слов;
- среднее количество знаков пунктуации на 1000 символов;
- частота 100 популярных биграмм букв на 1000 символов;
- частота служебных слов на 1000 символов;
- частота неопределенных местоимений на 1000 символов;
- частота коротких слов (менее 4 символов) на 1000 символов;
- частота длинных слов (более 7 символов) на 1000 символов;
- количество уникальных слов на 1000 символов;
- среднее число слов в предложении;
- количество грамматических ошибок на 1000 символов;
- количество предложений в тексте на 1000 символов;
- количество сложноподчиненных предложений на 1000 символов;
- доля сложноподчиненных предложений;
- количество вопросительных предложений на 1000 символов;
- количество восклицательных предложений на 1000 символов;
- доля вопросительных и восклицательных предложений;
- частота 100 популярных слов на 1000 символов;
- частота 100 популярных 2-грамм слов на 1000 символов;
- частота 100 популярных 3-грамм слов на 1000 символов;
- количество слов в семантическом ядре;
- наличие единства тематики в разных частях текста.

### 2.3.4 Расчет значений характеристик текста для наборов текстов известного происхождения

Для получения количественных значений выделенных характеристик текста были использованы расчеты, основанные на статистическом анализе, а именно оценке средних и среднеквадратичного отклонения значений характеристик.

Для каждой выборки были рассчитаны математическое ожидание  $M(X_k)$  и среднеквадратичное отклонение  $s(X_k)$  случайной величины  $X_k$  –  $k$ -й характеристики текста, которая является элементом множества исследуемых характеристик  $X$ :

$$X_k \in X; |X| = m; k = 1, \dots, m; m = 22.$$

Расчеты были проведены согласно алгоритму, представленному на рис. 2.4 в виде блок-схемы. Ниже приведено текстовое описание алгоритма расчетов статистических величин для одной характеристики  $X_k \in X$ .

Входными данными является выборка текстов, объединенных по происхождению (способу создания), представляющих собой множество  $T_x = \{ t \mid t - \text{текст выборки} \}$  мощности  $n$ .

Шаг 1. Для каждого текста  $t_i \in T_x$  вычисляется значение параметра  $X_{ki}$  ( $i = 1, \dots, n; k = 1, \dots, m$ ).

Шаг 2. Значения  $X_{ki}$  для  $i=1, \dots, (n-1)$  используются для вычислений математического ожидания  $M$  величины  $X_k$  и ее среднеквадратического отклонения  $s$ .

Шаг 3. Проверка отклонения:  $n$ -е значение величины  $X_k$  используется для проверки отклонения. Проверяется принадлежность  $X_{kn}$  отрезку числовой прямой  $(M(X_k) - s(X_k); M(X_k) + s(X_k))$ .

Шаг 4. В случае если проверка на шаге 3 дала отрицательный результат, рекомендуется заменить выборку частично или полностью для достижения положительного результата, далее перейти к шагу 1. Если

проверка дала положительный результат, то результатом (средним значением текстовой характеристики для исследуемой выборки) является математическое ожидание рассчитываемой случайной величины.

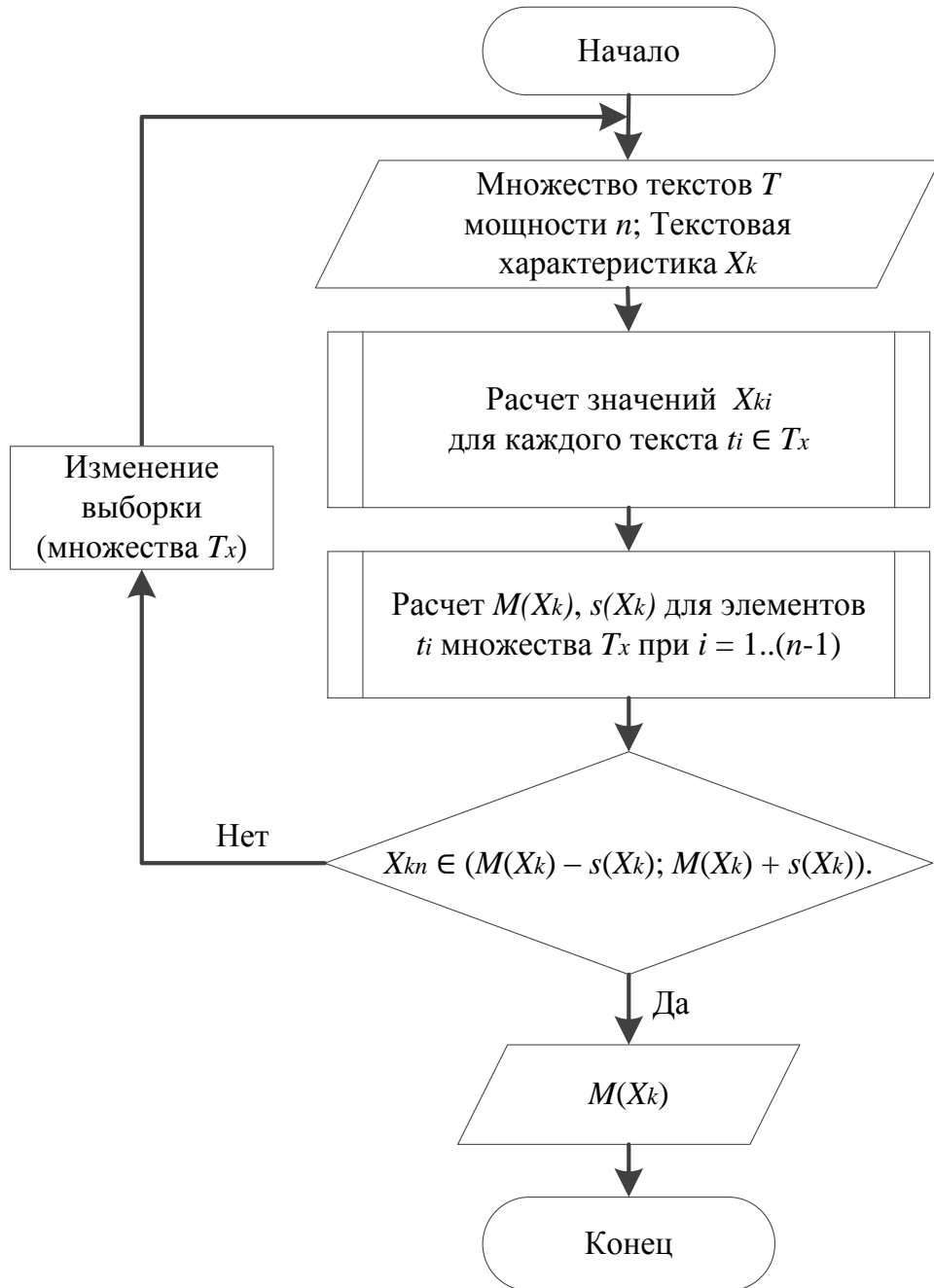


Рисунок 2.4 – Алгоритм вычисления статистических оценок значения характеристики  $X_k$  для текстовой выборки  $T_x$

Приведенный алгоритм для проведения расчетов был реализован в программном средстве Auth\_stat, свидетельство о государственной регистрации программы для ЭВМ приведено в приложении А. Данное программное средство проводит описанные выше вычисления, представляет полученные данные в виде таблицы, где столбцами являются выборки текстов (классы текстов), а строками – текстовые характеристики и их рассчитанные значения для каждой из выборки.

Рассчитанные с помощью программного средства Auth\_stat значения выделенных характеристик для трех выборок приведены в приложении Б. Графическое представление сравнения значений характеристик текстов разных выборок, нормированные в интервале  $[0; 1]$ , приведено в приложении В.

### **2.3.5 Проверка перечня характеристик на различительную способность**

По предложенной модели на следующем шаге необходимо оценить различительную способность выделенных характеристик, которая в данном случае выступает в качестве критерия информативности характеристик.

Для измерения различительной способности применяется мера, позволяющая оценить среднее расстояние между точками в пространстве [102]. Это позволяет определить критерии сравнения (характеристики текста), которые не принимают существенно различные значения для разных классов текстов или имеют сильное отклонение от среднего, а следовательно, не могут быть использованы в задачах текстовой атрибуции.

Для математических расчетов удаленности значений характеристики для текстов из разных выборок было использовано условие превосходства разности математических ожиданий для двух классов по модулю над суммой их среднеквадратических отклонений:

$$|M_1 - M_2| > s_1 + s_2,$$



где  $M_1$ ,  $M_2$  – математические ожидания величины значения текстовой характеристики для двух выборок текстов;  $s_1$ ,  $s_2$  – среднеквадратические отклонения величины значения текстовой характеристики для двух выборок текстов (индексы совпадают).

В приложении Г приведены результаты вычислений. В соответствии с полученными результатами был сделан вывод, что для исследуемых классов текстов различительной способностью не обладают следующие характеристики, которые были удалены из исследования:

- средняя длина слова;
- частота длинных слов на 1000 символов;
- доля сложноподчиненных предложений;
- доля восклицательных и вопросительных предложений.

### 2.3.6 Оценка взаимозависимости характеристик

Оценка корреляции рассчитанных значений позволяет выделить характеристики внутри одной выборки тестов, имеющие статистически значимые различия в значениях. Пары, имеющие сильную корреляционную зависимость, должны быть проанализированы и разделены: одна из характеристик удалена из набора. Это позволит снизить вычислительные затраты для расчета значений и классификации, а также увеличить различительную способность инварианта в целом.

Оценка корреляции значений текстовых характеристик внутри каждой выборки осуществлялась с помощью метода Пирсона (метод квадратов) [103]:

$$r_{xy} = \frac{\sum_{i=1}^k (d_{x_i} \cdot d_{y_i})}{\sqrt{\sum_{i=1}^k d_{x_i}^2 \cdot \sum_{i=1}^k d_{y_i}^2}},$$

где  $d_{xi}$ ,  $d_{yi}$  – отклонение  $i$ -го числового значения от среднего значения своего вариационного ряда;  $k$  – количество элементов вариационных рядов (количество текстов в наборе).

По итогам расчета коэффициента корреляции была обнаружена сильная корреляция ( $|r_{xy}| \geq 0,7$ ) у ряда пар характеристик. Были удалены следующие из них:

- частота неопределенных местоимений на 1000 символов;
- частота 100 популярных 3-грамм слов на 1000 символов;
- количество коротких слов на 1000 символов;
- количество восклицательных предложений на 1000 символов.

### 2.3.7 Полученные инварианты искусственных текстов

По результатам проведенных вычислений были сформированы три инварианта для выделенных классов текстов. Согласно обозначениям, приведенным в п. 2.1 работы, инвариант  $a_j \in A$  представляет собой массив упорядоченных пар:

$$a_j = (\langle x_1, z_{1j} \rangle, \dots, \langle x_i, z_{ij} \rangle, \dots, \langle x_m, z_{mj} \rangle),$$

где  $z_{ij}$  – значение характеристики текста  $x_i \in X$  для  $j$ -й выборки текста;  $i = 1, \dots, m; j = 1, \dots, n$ ;

Для удобства дальнейшего использования полученных инвариантов и представления их в виде координат точек в  $m$ -мерном пространстве было решено в дальнейшем описывать инварианты  $a_i \in A$  в виде вектора значений характеристик. Размерность таких векторов соответствует количеству отобранных характеристик текста, равному  $m$ :

$$a_j = (a_{j1}, a_{j2}, \dots, a_{jm}),$$

где  $a_{ji}$  – значение  $j$ -ой текстовой характеристики  $i$ -го инварианта,  $i = 1, \dots, n$ ,  $j = 1, \dots, m$ ;  $n$  – количество инвариантов (соответствует количеству классов текстов);  $m$  – количество используемых характеристик текста в инварианте.

Характеристики текста, составившие инварианты текстов, разделенных на классы по своему происхождению:

- среднее количество знаков пунктуации на 1000 знаков текста;
- частоты 100 популярных биграмм букв на 1000 символов;
- частота служебных слов на 1000 символов;
- количество уникальных слов на 1000 символов;
- среднее число слов в предложении;
- количество грамматических ошибок на 1000 символов;
- количество предложений на 1000 символов;
- количество сложноподчиненных предложений на 1000 символов;
- количество вопросительных предложений на 1000 символов;
- частота 100 популярных слов на 1000 символов;
- частота 100 популярных 2-грамм слов на 1000 символов;
- количество слов в семантическом ядре;
- наличие единства тематики в разных частях текста.

Ниже приведены полученные векторы численных значений характеристик текста:

$$a_1 = (31,742; 201,269; 34,691; 64,804; 9,113; 0,01; 109,812; 68,655; 1,414; 49,001; 9,1; 66,025; 1,7);$$

$$a_2 = (21,545; 184,36; 31,865; 66,155; 16,549; 1,002; 60,356; 31,658; 0,787; 37,104; 4,196; 75,268; 1,3);$$

$$a_3 = (29,035; 112,562; 25,702; 101,659; 9,987; 6,215; 100,2; 62,082; 1,358; 32,882; 3,554; 95,645; 0,6);$$

где  $a_1$  – инвариант класса естественных текстов;  $a_2$  – инвариант класса искусственных текстов, созданных на основе цепей Маркова;  $a_3$  – инвариант

класса искусственных текстов, сгенерированных с помощью синонимизации):

Таким образом, с помощью предложенной модели был получен набор характеристик текстов, обладающих различительной способностью в решении задачи определения происхождения текста, а именно определения, написан ли текст человеком или создан автоматически с помощью программного генератора. На основе проведенных расчетов средних были сформированы инварианты исследуемых классов текстов.

## 2.4 Выводы

В данной главе предложена модель процесса формирования инвариантов классов текстов, которая обобщает последовательность действий, используемые механизмы и ресурсы для данного процесса. Предложенная модель основана на использовании качественных и уточняющих их количественных характеристик текста.

Таким образом, при формировании набора исследователю основывается на лингвистических особенностях рассматриваемых классов текстов. Модель является универсальной, так как может быть применена при решении любой задачи, связанной с классификацией текстов.

На основе предложенной модели были выделены лингвистические особенности естественных текстов, в отличие от искусственных экземпляров. Согласно всем этапам процесса были проведены формирование перечня количественных характеристик текста, оценка их различительной способности и корреляции между собой. На основе проведенных расчетов средних были сформированы инварианты исследуемых классов текстов, которые могут быть использованы в решении задачи определения искусственно созданных текстов.

### **3 МЕТОД И ПРОГРАММНОЕ СРЕДСТВО ОПРЕДЕЛЕНИЯ ИСКУССТВЕННО СОЗДАНЫХ ТЕКСТОВ**

#### **3.1 Предлагаемый метод определения искусственно созданных текстов**

В соответствии с поставленной задачей диссертационного исследования выделены виды текстов, происхождение которых требуется определить. Ими является автоматически созданный (искусственный) информационный контент веб-ресурсов.

Таковыми текстами являются статьи, распространяемые в различных социальных сетях, посредством электронной почты и другими способами. В работе рассмотрены тексты на русском языке длиной от 1 200 до 5 200 символов, представляющие собой записи на форумах, в социальных сетях, соответствующие тематикам: общество, политика, силовые структуры, религиозные общины и другим, затрагивающим социально острые проблемы.

Под автоматизированным определением искусственного текста понимается принятие решения о происхождении произвольного входного текста: был ли текст написан человеком или текст был сгенерирован с помощью специального алгоритма. В исследованиях рассматривались алгоритмы синонимизации, метода, основанного на Марковских цепях, как наиболее распространенные алгоритмы, используемые для генерации интернет-контента [61].

Предлагается функциональная модель определения искусственно созданных текстов, представленная в виде контекстной диаграммы в нотации IDEF0 на рис. 3.1.



Рисунок 3.1 – Контекстная диаграмма определения искусственно созданных текстов

Исходя из представленной на рис. 3.1 диаграммы, можно заключить, что основой для принятия решения системе будут служить инварианты генераторов искусственных текстов.

Необходимо создать аппарат для математической оценки принадлежности входного текста к одному из выделенных классов текстов, а также для ее интерпретации.

В связи с изложенными фактами к методу определения искусственных текстов выдвигаются следующие требования:

- 1) метод принимает на вход текст на русском языке длиной от 1 200 символов;
- 2) метод должен формально определить принадлежность входного текста к одному из известных ей классов;
- 3) решение принимается на основе инвариантов классов естественных и искусственных текстов;

4) принятое решение должно иметь оценку точности отнесения текста к одному из классов.

Метод предназначен для определения происхождения текста, а именно определения, создан входной текст человеком или искусственно.

Область применения метода:

- поисковые системы;
- системы поиска спама, в том числе в составе сложных систем (sms, почтовые сервисы, социальные сети, форумы);
- системы определения плагиата;
- другие системы и сервисы, нуждающиеся в определении происхождения текста.

На вход метода подается текст на русском языке, длиной от 1 200 символов.

Выходными данными метода являются заключение о происхождении текста (текст написан человеком или сгенерирован автоматически) и точность, с которой выдано заключение.

Обозначения, используемые в описании метода:

$T$  – множество текстов, которые могут быть исследованы;

$X = \{x_1, x_2, \dots, x_m\}$  – множество исследуемых характеристик текста,  $|X| = m$ ;

$A = \{a_1, a_2, \dots, a_n\}$  – множество инвариантов классов текстов, разделенных по происхождению, то есть множество наборов усредненных значений характеристик текста  $x_j \in X$  для  $n$  классов текстов;

$a_i$  – инвариант  $i$ -го класса текста;  $a_i = (a_{i1}, a_{i2}, \dots, a_{im})$ , где  $a_{ij}$  – усредненное информативное значение  $j$ -й характеристики текста  $i$ -го инварианта,  $i = 1, \dots, n; j = 1, \dots, m$ .

$a'$  – набор рассчитанных значений характеристик текста некоторого входного текста  $t$ , происхождение которого требуется определить,  $t \in T$ ;

$a'_j$  – рассчитанное значение  $j$ -й текстовой характеристики входного текста,  $j = 1, \dots, m$ ;

$V(a', A)$  – мера оценки принадлежности входного текста к классу текстов с известным происхождением;

$D(a', a_i)$  – мера расстояния между входным текстом и  $i$ -м классом текстов, представляемая как мера расстояния между векторами  $a'$  и  $a_i$ ;

$l$  – пороговое значение расстояния между вектором значений текстовых характеристик входного текста  $a'$  и вектором-инвариантом  $i$ -го класса текста с известным происхождением  $a_i$  такое, что максимальное значение меры  $D(a', a_i)$  не должно превышать  $l$  при  $i=1, \dots, n$ ;

$R_A$  – точность решения о способе создания текста, которое принимается с помощью метода.

Метод определения искусственно созданных текстов представляет собой последовательность действий, связанных с анализом характеристик текста [104].

Шаг 1. Рассчитать числовые значения текстовых характеристик  $x_i \in X$  входного текста,  $i=1, \dots, n$ .

Шаг 2. Сформировать вектор  $a'$  как набор рассчитанных значений характеристик.  $a' = (a'_1, a'_2, \dots, a'_m)$ ;

Шаг 3. Рассчитать меры расстояний  $D(a', a_i)$  между векторами  $a'$  и  $a_i \in A$ ,  $i = 1, \dots, n$ .

Шаг 4. Рассчитать меру принадлежности входного текста к известным классам. Мера принадлежности определяется как выбор наименьшей из мер расстояний, рассчитанных на шаге 3:

$$V(a', A) = \min[D(a', a_i)]; i = 1, \dots, n.$$



Шаг 5. Сравнить меру принадлежности, полученную на шаге 4, с заданным пороговым значением  $l$ .

Шаг 6. Принять решение о происхождении текста. Принимается решение о том, что входной текст отнесен к  $i$ -му классу, если выполняются следующие соотношения:

$$V(a', A) \equiv D(a', a_i);$$

$$V(a', A) \leq l.$$

Принимается решение о том, что входной текст не может быть отнесен ни к одному из классов текстов известного происхождения, если выполняется соотношение:

$$V(a', A) > l.$$

Шаг 7. Рассчитать точность заключения о происхождении текста  $R_A$  по формуле:

$$R_A = 1 - \frac{V(a', A)}{D(a_x, a_y)},$$

где  $a_x, a_y \in A$  – два вектора-инварианта, наиболее приближенных (согласно мере расстояния) к  $a'$ .

### **Пояснения к методу определения искусственно созданных текстов**

1. Для расчета меры расстояния  $D(a', a_i)$  между векторами  $a'$  и  $a_i$  может быть использована любая метрика, позволяющая количественно оценить расстояние между двумя точками в  $k$ -мерном пространстве. Среди возможных метрик – общеизвестные евклидова метрика и метрика Махаланобиса [105–107]. Нужно отметить, что выбор метрики должен отвечать требованиям поставленной задачи.

Так как для решения задачи определения происхождения текста необходим единый масштаб получаемых значений меры  $D(a', a_i)$ , для ее

расчета автором предлагается использовать метрику (расстояние) Махаланобиса. Данная метрика обобщает понятие расстояния Евклида, учитывает корреляции между переменными и инвариантно к масштабу. Она широко используется в кластерном анализе и методах классификации.

Используя метрику Махаланобиса, меру расстояния между векторами  $a'$  и  $a_i$  можно представить следующим выражением:

$$D(a', a_i) = \sqrt{(a' - a_i)^T \cdot S^{-1} \cdot (a' - a_i)},$$

где  $S$  – объединенная ковариационная матрица.

При этом значения меры будут строго принадлежать интервалу  $D(a', a_i) \in [0; 1]$ . Значение величины  $l$  также должно лежать в этом интервале:  $l \in [0; 1]$ .

2. Пороговое значение  $l$  ограничено сверху величиной, равной половине меры расстояния между двумя наиболее приближенными векторами-инвариантами  $a_x, a_y \in A$  и может быть скорректировано в меньшую сторону на основе экспериментальных расчетов:

$$l \leq \frac{D(a_x; a_y)}{2}.$$

3. Значение величины точности заключения о происхождении текста (соответствии текста некоторому инварианту из множества  $A$ )  $R_A$  лежит в интервале  $[0; 1]$ . Ее значение тем выше, чем меньше расстояние между вектором значения характеристик входного текста с вектором-инвариантом, с которым он был соотнесен.

### **3.2 Программное средство фильтрации искусственно созданных текстов**

Было разработано специализированное программное средство «TextOrigin», предназначенное для автоматизации процесса обнаружения и последующей фильтрации искусственно сгенерированных текстов. В основе реализации разработанной системы используются сформированные в главе 2 инварианты естественных и искусственно созданных текстов. Процесс обнаружения искусственных текстов осуществляется согласно предложенному в п. 3.1 методу. Ниже представлены основные функции программы:

- анализ базы данных текстовой информации на предмет наличия искусственно созданных текстов;
- оповещение модератора в случае выявления в запросах на публикацию фактов использования искусственно созданных текстов;
- автоматическая фильтрация входных данных при достижении заданного порога, характеризующего текст как искусственный.

Для разработки программного средства была выбрана веб-ориентированная технология, позволяющая создавать кроссплатформенные решения, а также использовать для написания сервиса и его потребителя (т.е. клиента) разные языки программирования. Кроме того, базовые возможности программного средства были реализованы в виде встраиваемых модулей для популярных систем управления контентом. Таким образом, автором был предоставлен удобный функционал для интеграции разработанной системы в действующие веб-ресурсы, обладающие потребностью осуществлять анализ входных данных на критерий искусственности их происхождения. Практическая ценность разработанного средства заключается в возможности его использования в качестве фильтра для:

- модулей загрузки «авторских» новостей в СМИ;

- модулей приема заявок / обращений онлайн;
- модулей загрузки постов в сообществах социальных сетей.

### 3.2.1 Требования к программному средству

Поставленные в настоящем исследовании задачи предполагают разработку программного средства, позволяющего анализировать входную информацию на предмет выявления фактов ее искусственной генерации. Кроме того, необходимо предусмотреть функционал фильтра, позволяющего в автоматическом режиме производить «чистку» больших массивов подобной информации.

В главе 1 было установлено, что площадкой для внедрения подобного средства может выступать множество интернет-ресурсов, предоставляющих возможность своим пользователям самостоятельно загружать контент в БД: от сайтов с формой обратной связи до чатов, социальных сетей, форумов, медиа-порталов и т.д.

В связи с этим были выдвинуты требования к разрабатываемой системе:

- возможность многопоточного режима работы (одновременно должны обрабатываться несколько запросов от разных источников). Данное требование обусловлено особенностями рассматриваемых интернет-ресурсов;
- возможность обработки файлов-источников популярных текстовых форматов: «txt», «doc», «docx», «odt», «rtf»);
- возможность структуризации входных данных (обработка входных данных, использующих языки разметки и разделение информации на метаданные, описывающие структуру документа);
- обработка русскоязычных текстов;
- предоставление современных графических инструментов представления результатов анализа и фильтрации;

– кроссплатформенность, кроссбраузерность, а также адаптивный дизайн в целях обеспечения возможности модератору управлять работой системы с любого типа устройств.

### **3.2.2 Структура программного средства**

Функционал разработанного программного комплекса реализован в виде трех взаимосвязанных подсистем, каждая из которых состоит из совокупности модулей. На рис. 3.2 приведена структура системы.

Ниже представлено краткое описание задач, выполняемых представленными на рис. 3.2 компонентами программного средства.

#### **Аналитическая подсистема**

Основополагающим элементом «TextOrigin» является аналитическая подсистема, совокупность модулей которой производит весь процесс подготовки, обработки и анализа входных текстов. В составе подсистемы выделяются следующие модули:

##### *Модуль экранирования*

Прежде чем начать обработку поступающего контента, необходимо предусмотреть механизмы защиты системы от возможных злоумышленных действий. Один из распространенных способов атаки на сайты и сервисы, работающие с БД, основан на внедрении в форму ввода произвольного SQL-кода. Эксплуатация злоумышленниками подобных уязвимостей может привести к необратимым последствиям вплоть до уничтожения всей БД и потери контроля над системой в целом [108]. Данный модуль отвечает за экранирование всех потенциально опасных символов и команд, не позволяя помещать в SQL-запрос управляющие структуры и идентификаторы, введенные пользователем.

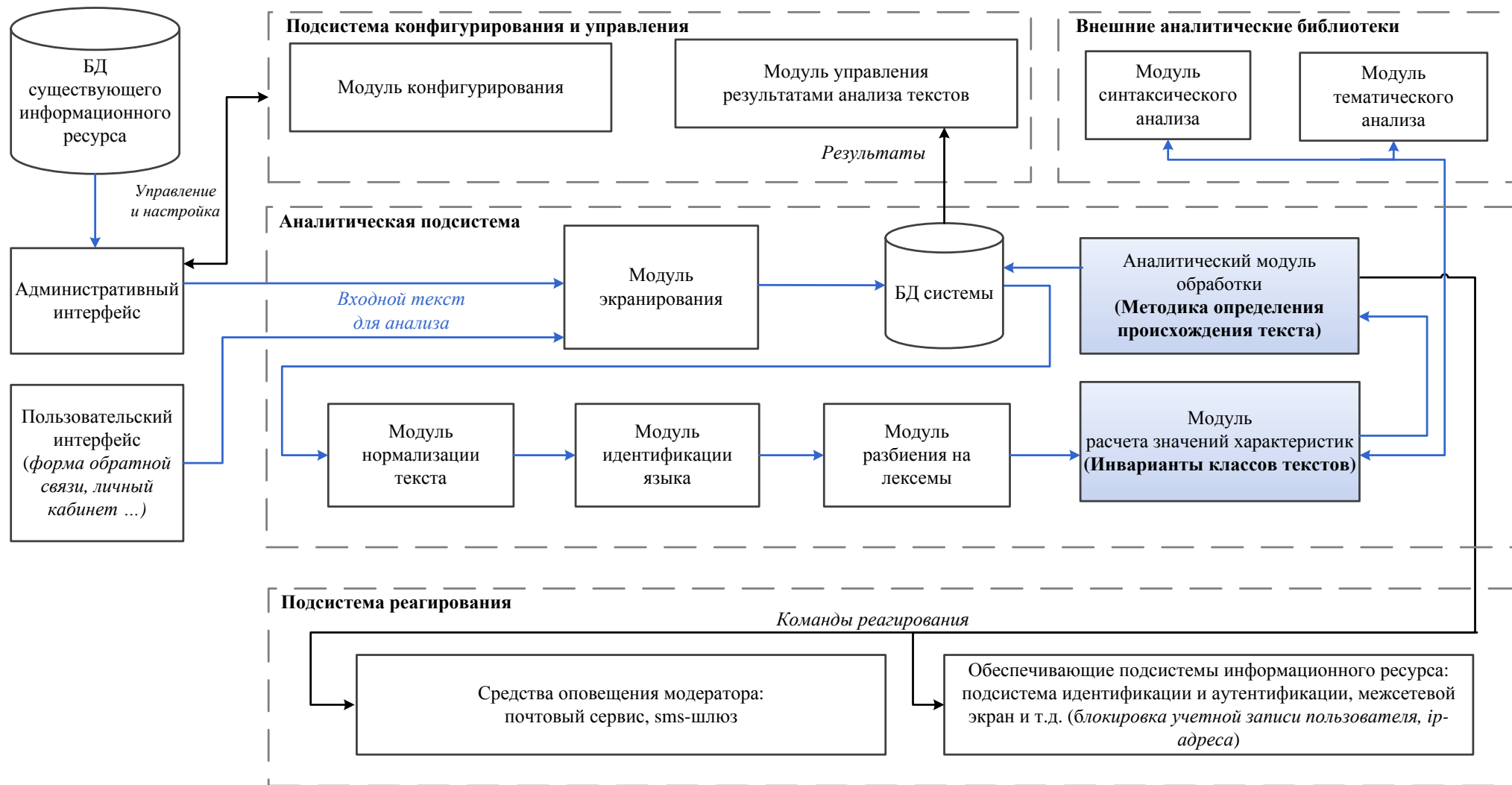


Рисунок 3.2 – Структура программного средства, позволяющего определять происхождение входного текста

### *Модуль нормализации текста*

Следующим этапом является предварительная обработка текста, представляющая собой автоматизированный процесс подготовки содержимого к дальнейшему анализу путем его приведения к формату входных данных предложенного автором метода. Следует отметить, что природа происхождения исследуемого текста может быть самой разнообразной. Например, загружаемые сведения могут быть скопированы с интернет-ресурсов, в том числе в формате HTML, распознаны с печатного источника или набраны автором в текстовом процессоре. Этим обуславливается и многообразие используемых подходов и особенностей, которые необходимо учитывать на этапе предварительной подготовки контента. Тем не менее, основными этапами нормализации текста являются:

1) изменение кодировки на универсальную или подходящую под конкретную задачу. Разработанное программное средство преобразует текст в соответствии со стандартом Юникод (UTF-8) [109];

2) выделение текста, то есть удаление возможных дополнительных материалов: изображений, таблиц, тегов, служебных символов или комментариев;

3) корректировка ошибок: удаление «лишних», не относящихся к тексту символов (например, символ перевода строки, двойные тире, двойные пробелы и т.д.);

4) замена равнозначных символов языка для упрощения дальнейшего анализа, например «ё» на «е»;

5) унификация знаков препинания: кавычек, тире, апострофов.

### *Модуль идентификации языка*

Разработанное средство позволяет анализировать происхождение исключительно русскоязычных текстов. На практике загружаемые тексты зачастую являются мультязычными (содержат заголовки, словосочетания,

а порой и цитаты большой размерности на иностранном языке). Данный модуль исключает из входного текста подобные вхождения и в дальнейшем уведомляет об этом администратора.

#### *Модуль разбиения на лексемы*

Под лексемой понимается словарная единица языка, представляющая текстообразующий компонент. Так как метод определения искусственных текстов, предложенная в п. 3.1 диссертационной работы, основывается на анализе количественных значений характеристик текста, связанных с текстообразующими компонентами, необходимо провести предварительное разбиение входного текста на лексемы, представив, таким образом, исследуемый текст в виде упорядоченного массива слов. Данный подход позволяет эффективно исследовать характеристики на символьном и лексическом уровнях. Современные языки программирования представляют широкий спектр стандартных возможностей (библиотек) для работы со строками. В числе этих возможностей – разбиение на подстроки, что позволяет успешно и с минимальными вычислительными затратами осуществить работу данного модуля. На выходе данный текст представляется как упорядоченный массив слов, предназначенный для дальнейшей обработки модулем расчета значений характеристик.

#### *Модуль расчета значений характеристик*

Расчет численных значений характеристик входного текста – важнейший этап работы программного средства, так как его итогом является набор значений, который будет использован для принятия решения о происхождении текста. Согласно обозначениям предложенного в п. 3.1 метода модуль формирует набор значений  $a'$ . На вход модуля подается исследуемый текст, представленный в виде массива слов, а также текст в виде строки. На выходе – значения характеристик данного текста.

Расчет значений производится в 2 этапа: первый подразумевает работу с текстом в виде упорядоченного массива лексем, а второй – с



нормализованным текстом, представленным в виде единой строки. Это обусловлено особенностями вычисления значений определенных характеристик. На основе текста, представленного в виде массива лексем, с использованием библиотек языка программирования РНР для работы со строками [110, 111] производится расчет следующих характеристик:

- частоты 100 популярных биграмм букв;
- частота служебных слов;
- количество уникальных слов;
- частота 100 популярных слов;
- частота 100 популярных 2-грамм слов;
- количество слов в семантическом ядре.

Анализ синтаксических характеристик выполняется средствами интерфейса прикладного программирования внешней аналитической библиотеки [112], позволяющей выполнять процедуру синтаксического разбора, в результате чего осуществляется расчет значений характеристик текста, представленного в виде строки:

- среднее количество знаков пунктуации в предложении;
- среднее число слов в предложении;
- количество грамматических ошибок;
- количество предложений;
- количество сложноподчиненных предложений;
- количество вопросительных предложений.

Для исследования единственной тематической характеристики – наличия единства тематики в разных частях текста – используется внешний модуль тематического анализа [113], представляющий собой автоматический рубрикатор с собственной БД, предназначенной для коротких отрывков текстов. Данный рубрикатор обладает возможностью использования интерфейса прикладного программирования.

### *Аналитический модуль обработки*

В данном модуле реализованы расчеты на основе метода идентификации искусственно созданных текстов, предложенного автором в п. 3.1. Для расчета меры расстояния между векторами используется метрика Махаланобиса. Расчеты производятся средствами штатных математических функций [114, 115]. На вход модуля подаются сформированные в главе 2 инварианты классов естественных и искусственных текстов, а также набор численных значений характеристик входного текста. На выходе модуля – заключение о происхождении текста с указанием точности отнесения тесту к одному из классов.

### **Подсистема реагирования**

В программном средстве «TextOrigin» предоставлен богатый набор инструментов для осуществления различных действий в случае обнаружения признаков искусственно созданного текста. Система позволяет настроить реакцию в виде:

- оповещения модератора посредством электронного письма на e-mail;
- оповещения модератора посредством SMS сообщения (используя сторонний шлюз);
- отправки сообщения на сервер мониторинга посредством протокола SNMP [116];
- выполнения произвольных скриптов с целью запуска или изменения настроек системного ПО. Например, администратор обладает возможностью настроить автоматизированный запуск скрипта, добавляющий адрес источника, загрузившего искусственный текст, в запрещающий список системного межсетевого экрана.

## Подсистема конфигурирования и управления

Модуль конфигурирования позволяет администратору системы производить настройку работы аналитического модуля, а также редактировать параметры взаимодействия с задействованными внешними информационными системами и сервисами. К числу редактируемых параметров относятся:

- идентифицируемые методы генерации;
- значение порогового расстояния между вектором текстовых характеристик входного контента и вектора-инварианта класса текста с известным происхождением;
- правила экранирования подозрительных подстрок;
- правила нормализации текста (массив выполняемых автозамен);
- правила реагирования на факт выявления искусственных текстов.

Существует возможность оповещения модератора путем настройки отправки сообщений через SMS-шлюз, передачу SNMP-ловушек серверу мониторинга, отправку уведомлений по e-mail. Кроме того, администратор может установить правила выполнения произвольных скриптов с целью запуска или изменения настроек системного ПО;

- параметры подключения внешних библиотек синтаксического и тематического анализа;
- параметры использования внешних ресурсов (подключение к БД, список анализируемых таблиц);
- параметры учетных записей пользователей и разграничения доступа.

Второй модуль данной подсистемы предоставляет функционал для управления результатами анализа. Существует возможность просмотра, визуализации в виде графиков, а также экспорта статистики и результатов анализа по всем проведенным ранее исследованиям текстов.

## Описание работы системы

Программное средство может обрабатывать входные данные, представленные множеством широко используемых текстовых форматов. Примерами файлов-источников могут выступать: неформатированные данные (формат «txt»), форматированные («doc», «docx», «odt», «rtf»). Отличительной особенностью системы является возможность обработки документов, использующие языки разметки, такие как html-тэги и xml-сущности. Структуризация входных данных осуществляется модулем нормализации текста.

В БД системы хранятся инварианты естественных и искусственных русскоязычных текстов. На рисунке 3.3 представлен пользовательский интерфейс сервиса, предоставляющего возможности загрузки новости на одном из популярных медиа-порталов Томской области.

Есть новость

Все поля являются обязательными для заполнения, кроме специально указанных.

Дата события  
02.05.2016 12:43:10

Текст новости  
Попечительский совет Международной Федерации RoboCup предложил Томской области стать одним из организаторов чемпионата Super Regional RoboCup Asia-Pacific. Решение попечительского совета было принято на заседании в Лейпциге по итогам конкурса на право проведения мирового финала RoboCup в 2018 году. Делегацию Томской области, принимавшей участие в конкурсе, возглавил заместитель губернатора по научно-образовательному комплексу и инновационной политике Михаил

Как с Вами связаться

Загрузить фото/аудио/картинку (Файл не может быть больше 5 мегабайт)

Список файлов	Вес	Статус
robocup.jpg	39 KB	100% <span style="color: green;">✔</span>
robocup2.jpg	39 KB	100% <span style="color: green;">✔</span>

Загружено 2/2 файлов 78 KB 100%

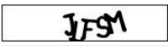
Введите код с картинки  
  
 JLDSM

Рисунок 3.3 – Скриншот результата работы системы, интегрированной в новостной портал. Сервис для пользователя «Есть новость»

Модератор в онлайн-режиме может получать информацию о результатах проведенных анализов на выявление искусственных текстов. Пример консоли модератора приведен на рисунке 3.4.

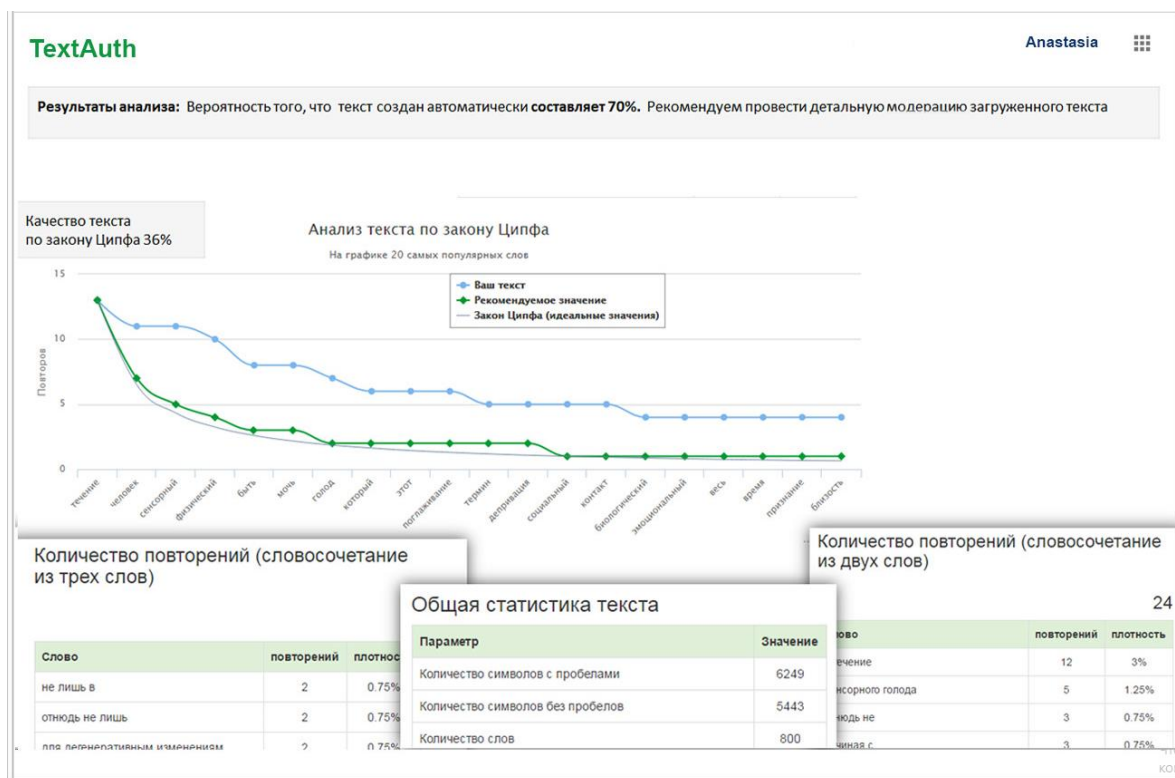


Рисунок 3.4 – Консоль модератора в системе «TextOrigin».

Для разработки программного средства были использованы языки программирования PHP [117] и Perl [118]. В качестве фреймворка был использован Bitrix Framework – технологическое ядро (платформа) для создания и управления проектами в составе продукта «1С-Битрикс: система управления сайтом». Основными преимуществами выбранной платформы являются принцип модульности и высокий уровень защищенности от взлома, подтвержденный результатами независимых аудитов [119].

Возможность многопоточного режима обработки множества запросов обеспечена применением веб-кластера «1С Битрикс» (рис. 3.5). Для

балансировки нагрузки между нодами используется модуль ngx\_http\_upstream http-сервера Nginx [120].

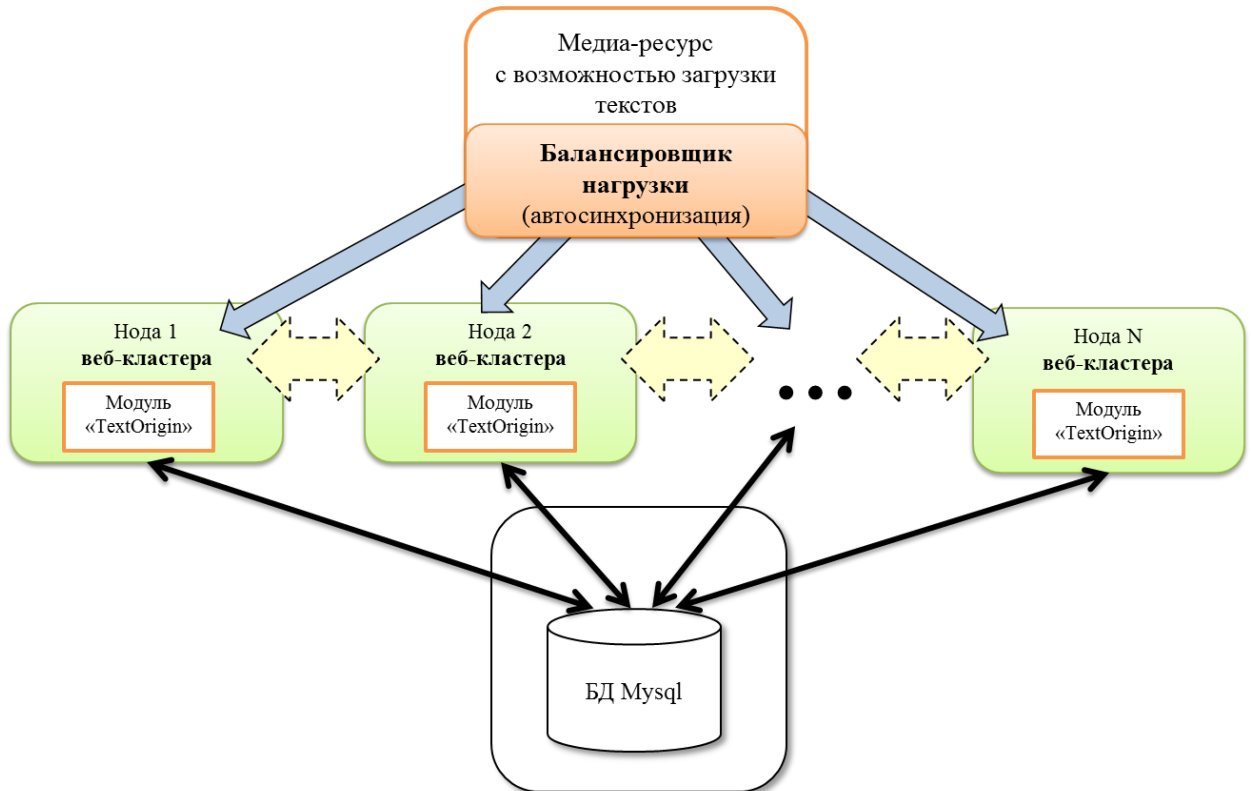


Рисунок 3.5 – Балансировка нагрузки, реализованная посредством применения модуля «Веб-кластер» продукта «1С Битрикс: Управление сайтом».

Кроссплатформенность и кроссбраузерность обеспечивается использованием современных web-технологий. Применение CSS-фреймворка Bootstrap [121] позволило реализовать дружественный интерфейс с адаптивным дизайном, позволяющим модератору управлять работой системы с любого типа устройств: смартфона, планшета или персонального компьютера.

### 3.3 Тестирование метода и программного средства

Для оценки эффективности предложенного в п. 3.1 метода определения искусственно созданных текстов были проведены экспериментальные расчеты меры принадлежности некоторых входных текстов известного происхождения к исследуемым классам текстов.

Целью проведения серии экспериментов является получение объективных сведений об эффективности предложенного автором метода определения искусственно созданных текстов.

В экспериментах были использованы тексты различного происхождения, собранные автором работы в популярных социальных сетях, а также предоставленные компаниями-партнерами для апробации. Объем каждой статьи составлял от 1 200 до 5 200 символов. Общий объем исходной выборки естественных текстов составил 1,965 млн символов. Статьи отбирались по тематикам, определенным в п. 3.1: общество, политика, финансы, власть, армия, силовые структуры, наука и техника и смежные с ними.

Естественные тексты подверглись действию автоматических генераторов на основе синонимизации и метода Марковских цепей для того, чтобы создать искусственные экземпляры. Для дополнительной оценки влияния объема словаря синонимов и, как следствие, уникальности искусственного текста на результат определения происхождения текста были использованы 2 словаря: с 700 тыс. синонимов и с 130 тыс. синонимов. Среднее значение уникальности искусственных текстов, созданных с помощью словаря с меньшим объемом, рассчитанное с помощью алгоритма шинглов, составило 36,5%, тогда как для словаря с большим объемом это значение составило 69,1%.

Таким образом, в эксперименте использовано по 1 000 текстов, объединенных одним из способов создания:

– естественные тексты, написанные человеком или несколькими людьми;

– искусственные тексты, созданные с помощью метода Марковских цепей;

– искусственные тексты, созданные с помощью синонимизации с помощью словаря из 700 тыс. синонимов;

– искусственные тексты, созданные с помощью синонимизации с помощью словаря из 130 тыс. синонимов.

Использовались следующие обозначения, соответствующие описанным выше:

$a'$  – набор рассчитанных значений характеристик входного текста;

$a_1$  – инвариант класса естественных текстов;

$a_2$  – инвариант класса искусственных текстов, созданных на основе цепей Маркова;

$a_3$  – инвариант класса искусственных текстов, сгенерированных с помощью синонимизации;

$A$  – множество инвариантов.  $A = \{a_1, a_2, a_3\}$ .

Результаты расчетов меры расстояния  $D(a', a_i)$  до каждого класса и результирующая мера принадлежности  $V(a', A)$  входного текста приведены в табл. 3.1–3.3. Данные приведены для первых 10 текстов каждой из выборок. Приведенные числовые значения и решение о происхождении входного текста были приняты разработанным программным средством на основе приведенного в п. 3.1 метода определения искусственных текстов.



Таблица 3.1 – Значения мер близости, меры принадлежности для естественных входных текстов

$a'$ (п/п)	$D(a', a_1)$	$D(a', a_2)$	$D(a', a_3)$	$V(a', A) \equiv D(a', a_i)$	Заключение о происхождении текста	$R_A$
На входе – естественные тексты (созданные человеком или несколькими людьми)						
1	0,121	0,311	0,365	$a_i = a_1$	Естественный	0,85
2	0,058	0,480	0,302	$a_i = a_1$	Естественный	0,87
3	0,214	0,501	0,451	$a_i = a_1$	Естественный	0,54
4	0,125	0,398	0,370	$a_i = a_1$	Естественный	0,73
5	0,174	0,413	0,450	$a_i = a_1$	Естественный	0,78
6	0,270	0,407	0,398	$a_i = a_1$	Естественный	0,54
7	0,116	0,457	0,376	$a_i = a_1$	Естественный	0,69
8	0,041	0,579	0,301	$a_i = a_1$	Естественный	0,91
9	0,149	0,325	0,384	$a_i = a_1$	Естественный	0,81
10	0,240	0,306	0,415	$a_i = a_1$	Естественный	0,70

Таблица 3.2 – Значения мер близости, меры принадлежности входных текстов для искусственных текстов, созданных с помощью метода Марковских цепей

$a'$ (п/п)	$D(a', a_1)$	$D(a', a_2)$	$D(a', a_3)$	$V(a', A) \equiv D(a', a_i)$	Заключение о происхождении текста	$R_A$
На входе – искусственные тексты, созданные с помощью метода цепей Маркова						
1	0,351	0,203	0,071	$a_i = a_3$	Искусственный, (Марк. цеп.)	0,85
2	0,490	0,341	0,105	$a_i = a_3$	Искусственный, (Марк. цеп.)	0,87
3	0,391	0,356	0,160	$a_i = a_3$	Искусственный, (Марк. цеп.)	0,54
4	0,405	0,327	0,101	$a_i = a_3$	Искусственный, (Марк. цеп.)	0,73
5	0,480	0,406	0,126	$a_i = a_3$	Искусственный, (Марк. цеп.)	0,78
6	0,510	0,388	0,055	$a_i = a_3$	Искусственный, (Марк. цеп.)	0,54
7	0,368	0,365	0,108	$a_i = a_3$	Искусственный, (Марк. цеп.)	0,69
8	0,371	0,348	0,040	$a_i = a_3$	Искусственный, (Марк. цеп.)	0,91
9	0,446	0,420	0,166	$a_i = a_3$	Искусственный, (Марк. цеп.)	0,81
10	0,409	0,387	0,089	$a_i = a_3$	Искусственный, (Марк. цеп.)	0,70

Таблица 3.3 – Значения мер близости, меры принадлежности входных текстов для искусственных текстов, созданных с помощью синонимизации

$a'$ (П/П)	$D(a', a_1)$	$D(a', a_2)$	$D(a', a_3)$	$V(a', A) \equiv D(a', a_i)$	Заключение о происхождении текста	$R_A$
На входе – искусственные тексты, созданные с помощью синонимизации на базе словаря из 700 тысяч синонимов						
1	0,351	0,203	0,071	$a_i = a_2$	Искусственный, (Синоним.)	0,85
2	0,490	0,341	0,105	$a_i = a_2$	Искусственный, (Синоним.)	0,87
3	0,391	0,356	0,160	$a_i = a_2$	Искусственный, (Синоним.)	0,54
4	0,405	0,327	0,101	$a_i = a_2$	Искусственный, (Синоним.)	0,73
5	0,480	0,406	0,126	$a_i = a_2$	Искусственный, (Синоним.)	0,78
6	0,510	0,388	0,055	$a_i = a_2$	Искусственный, (Синоним.)	0,54
7	0,368	0,365	0,108	$a_i = a_2$	Искусственный, (Синоним.)	0,69
8	0,371	0,348	0,040	$a_i = a_2$	Искусственный, (Синоним.)	0,91
9	0,446	0,420	0,166	$a_i = a_2$	Искусственный, (Синоним.)	0,81
10	0,409	0,387	0,089	$a_i = a_2$	Искусственный, (Синоним.)	0,70
На входе – искусственные тексты, созданные с помощью синонимизации на базе словаря из 130 тысяч синонимов						
11	0,218	0,307	0,261	$a_i = a_1$	Естественный	0,53
12	0,240	0,201	0,341	$a_i = a_2$	Искусственный, (Синоним.)	0,75
13	0,298	0,231	0,480	$a_i = a_2$	Искусственный, (Синоним.)	0,71
14	0,058	0,121	0,473	$a_i = a_1$	Естественный	0,93
15	0,220	0,078	0,403	$a_i = a_2$	Искусственный, (Синоним.)	0,90
16	0,306	0,101	0,361	$a_i = a_2$	Искусственный, (Синоним.)	0,87
17	0,380	0,356	0,423	$a_i = a_2$	Искусственный, (Синоним.)	0,55
18	0,160	0,114	0,260	$a_i = a_2$	Искусственный, (Синоним.)	0,85
19	0,405	0,330	0,343	$a_i \notin A$	–	–
20	0,280	0,197	0,401	$a_i = a_2$	Искусственный, (Синоним.)	0,75

В табл. 3.4 приведены показатели ошибок 1 и 2 рода на основе полученных результатов проведенных вычислений для полных выборок. Под ошибками 1 рода понимаются случаи, когда естественный текст был принят за искусственный (ложноположительное событие, или «ложная тревога» для пользователя системы). Ошибки 2 рода указывают на случаи, когда искусственный текст не был распознан системой и был принят за естественный экземпляр (ложноотрицательное событие, или «пропуск события»).

Таблица 3.4 – Показатели ошибок 1 и 2 рода

Показатель	Ошибки 1-го рода	Ошибки 2-го рода
Определение текста, созданного с помощью синонимизации (любительский словарь – среднее значение уникальности 36 %)	3,2 %	6,1%
Определение текста, созданного с помощью синонимизации (дополненный словарь – среднее значение уникальности 69 %)		2,8%
Определение текста, созданного с помощью алгоритма на основе цепей Маркова		2,6%

Показатель ошибок 1-го рода по определению искусственных текстов в целом (для текстов уникальностью выше 50 %) составил не более 4 %, ошибок 2 рода – не более 3 %. Высокие результаты определения искусственно созданных текстов показывают эффективность применения разработанного автором метода, а также инвариантов искусственных текстов.

По показателю ошибок 1 рода разработанный метод эффективнее методики Павлова А.С. [14] обнаружения искусственных текстов, представляющих собой поисковый спам (5,8 %). По показателям ошибок 2-го

рода разработанный метод имеет меньшую эффективность (указанная методика имеет показатель 1,7 %), однако достигнутый результат считается умеренным для систем текстовой атрибуции. Также следует отметить, что метода показывает большую эффективность при определении искусственных текстов, которые обладают уникальностью выше 50%.

### **3.4 Выводы**

В данной главе приведено описание предложенного оригинального метода определения искусственно созданных текстов, отличающегося использованием меры принадлежности. Были рассмотрены метод синонимизации и метод Марковских цепей в качестве генераторов текста как наиболее распространенные для генерации интернет-контента.

В соответствии с заданной проблемой и предметной областью были исследованы тексты на русском языке длиной выше 1 200 символов, представляющие собой записи на форумах, в социальных сетях и соответствующие тематикам: общество, политика, силовые структуры, религиозные общины и другим, затрагивающим социально острые проблемы.

Предложенный автором метод предназначен для определения способа создания текста и представляет собой последовательность действий, связанных с анализом характеристик текста.

На основе предложенного метода было разработано специализированное программное средство «TextOrigin», предназначенное для автоматизации процесса обнаружения и последующей фильтрации искусственно сгенерированных текстов. В основе реализации разработанной системы используются сформированные в главе 2 инварианты естественных и искусственно созданных текстов. Процесс обнаружения искусственных

текстов, а именно – проведение соответствующих расчетов и принятие решения, осуществляется согласно разработанному методу.

На основе произведенных экспериментальных вычислений с участием выборок текстов известного происхождения были определены показатели эффективности: ошибки 1 рода составили не более 4 %, ошибки 2 рода – не более 3 %. По показателю ошибок 1 рода разработанный метод эффективнее методики [14] обнаружения искусственных текстов, представляющих собой поисковый спам, предложенной Павловым А.С., данная методика имеет показатель ошибок 5,8 %. По показателям ошибок 2-го рода разработанный метод имеет меньшую эффективность (указанная методика имеет показатель ошибок 1,7 %), однако достигнутый результат считается умеренным для систем текстовой атрибуции. Также следует отметить, что метод показывает большую эффективность при определении искусственных текстов, которые обладают уникальностью выше 50%.

## ЗАКЛЮЧЕНИЕ

В работе приведены результаты комплексного исследования по решению задачи определения искусственно созданных текстов. Данная задача является междисциплинарной и охватывает научные знания прикладной лингвистики, текстовой атрибуции, стилометрии, статистического анализа.

В первой главе автором проведен обзор существующих методов и алгоритмов определения искусственных текстов. Среди них решения для выявления поискового спама, а также машинного перевода с одного естественного языка на другой. Существующие решения по определению неестественных текстов из-за наложенных ограничений и узкой направленности не могут быть использованы в задаче определения текстов, созданных с помощью синонимизаторов или других средств автоматической генерации. Это подтверждает актуальность работы и необходимость разработки метода определения искусственно созданных текстов.

Во второй главе предложен алгоритм формирования инвариантов классов текстов, которая обобщает последовательность действий, используемые механизмы и ресурсы для данного процесса. Предложенный алгоритм основан на использовании качественных и уточняющих их количественных характеристик текста. Таким образом, при формировании набора исследователь основывается на лингвистических особенностях рассматриваемых классов текстов.

На основе предложенной модели были выделены лингвистические особенности естественных текстов, в отличие от искусственных экземпляров. Согласно всем этапам процесса были проведены формирование перечня количественных характеристик текста, оценка их различительной способности и корреляции между собой. На основе проведенных расчетов

средних были сформированы инварианты исследуемых классов текстов, которые могут быть использованы в решении задачи определения искусственно созданных текстов.

В третьей главе предложен оригинальный метод определения искусственно созданных текстов, отличающаяся использованием меры принадлежности. Были рассмотрены такие методы генерации текста, как синонимизация и метод Марковских цепей, так как они наиболее распространены для генерации интернет-контента. Метод предназначен для определения происхождения текста и представляет собой последовательность действий, связанных с анализом характеристик текста.

На основе предложенного метода было разработано специализированное программное средство «TextOrigin», предназначенное для автоматизации процесса обнаружения и последующей фильтрации искусственно сгенерированных текстов. В основе реализации разработанной системы используются сформированные в главе 2 инварианты естественных и искусственно созданных текстов. Процесс обнаружения искусственных текстов осуществляется согласно разработанному методу.

По итогам произведенных экспериментальных вычислений с участием выборок текстов известного происхождения были определены показатели эффективности: ошибки 1-го рода составили не более 4 %, ошибки 2-го рода – не более 3 %. Данные результаты превосходят аналогичные методы для определения поискового спама и являются допустимыми в решении задачи текстовой атрибуции.

Результаты диссертационного исследования были внедрены в деятельность ООО «Агентство медиарешений», где показали свою эффективность в решении задачи определения искусственно созданных текстов. Также полученные наработки были внедрены в учебную деятельность ТУСУРа по дисциплинам «Дискретная математика», «Теория

вероятностей и математическая статистика». Факты внедрения результатов диссертационной работы подтверждаются соответствующими актами, приведенными в приложении Д.

По итогам проведенного диссертационного исследования была достигнута цель работы – повышена точность определения искусственно созданных текстов за счет того, что впервые были исследованы тексты, представляющие собой информационный контент. Отличием данных текстов является их предназначение – быть прочитанными пользователями, в отличие от поискового спама, который в первую очередь направлен на взаимодействие с поисковой машиной и зачастую не предназначен для прочтения пользователем. Значимость решения данной задачи повышается на фоне роста популярности интернет-ресурсов в последнее десятилетие: они становятся платформой для высказывания собственного мнения, агитации, рекламы, а также широко используются в качестве средств массовой информации.



## СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Использование интернета в террористических целях [Электронный ресурс] / Организация объединенных наций, Управление ООН по наркотикам и преступности. – Нью-Йорк, 2013. – 148 с. – Электрон. версия печатн. публ. – URL: [https://www.unodc.org/documents/terrorism/Publications/Use\\_of\\_Internet\\_for\\_Terrorist\\_Purposes/Use\\_of\\_the\\_internet\\_for\\_terrorist\\_purposes\\_Russian.pdf](https://www.unodc.org/documents/terrorism/Publications/Use_of_Internet_for_Terrorist_Purposes/Use_of_the_internet_for_terrorist_purposes_Russian.pdf) (дата обращения: 26.05.2016).
2. Хмелёв Д.В. Распознавание автора текста с использованием цепей А.А. Маркова // Вестник Московского государственного университета. Сер. 9, филология. – 2000. – № 2. – С. 115–126.
3. Khmelev D. Disputed Authorship Resolution Using Relative Empirical Entropy For Markov Chain of Letters in a Text [Electronic resource] / D. Khmelev // Proceedings of the 4th conference of the International Quantitative Linguistics Association, August 24–26, 2000, Prague / ed. by R.H. Baayen. – The electronic version of the printing publication. – URL: <http://www.philol.msu.ru/~lex/khmelev/proceedings/qualico2000a.pdf> (access date: 22.08.2016).
4. Киреев К. Штапомер – описание работы программы [Электронный ресурс] / К. Киреев. – Электрон. дан. – [Б. м.], 2011. – URL: <http://shtampomer.narod.ru/manual.html> (дата обращения: 15.08.2016).
5. Шевелев О.Г. Анализ частоты встречаемости различных длин предложений в литературном тексте как возможной характеристики авторского стиля с помощью самоорганизующихся карт кохонена / О.Г. Шевелев // Нейроинформатики и ее приложения : материалы XII Всероссийского семинара, 1–3 октября 2004 г., Красноярск / Ин-т вычислительного моделирования СО РАН. – Красноярск, 2004. С. 177–178
6. Шевелев О.Г. Классификация текстов с помощью деревьев решений и сетей прямого распространения / О.Г. Шевелев, А.В. Петраков //

Вестник Томского государственного университета. – 2006. – № 290. – С. 300–307.

7. Романов А.С. Методика и программный комплекс для идентификации автора неизвестного текста : автореф. дис. ... канд. техн. наук / А.С. Романов. – Томск, 2010. – 26 с.

8. Corney M. Identifying the Authors of Suspect Email [Electronic resource] / M. Corney, A. Anderson, G. Mohay, O. de Vel // QUT ePrints. – Electronic data. – Brisbane, [s. a.]. – URL: <http://eprints.qut.edu.au/8021/1/CompSecurityPaper.pdf> (access date: 22.08.2016).

9. Chakraborty T. Authorship Identification in Bengali Literature: a Comparative Analysis [Electronic resource] / T. Chakraborty // Researchgate.net. – Electronic data. – Berlin, San Francisco, 2012. – URL: [https://www.researchgate.net/publication/230764356\\_Authorship\\_Identification\\_in\\_Bengali\\_Literature\\_a\\_Comparative\\_Analysis](https://www.researchgate.net/publication/230764356_Authorship_Identification_in_Bengali_Literature_a_Comparative_Analysis) (access date: 22.08.2016).

10. Stylometry for E-mail Author Identification and Authentication [Electronic resource] / K. Calix [et al.] // Proceedings of CSIS Research Day, Pace University. – 2008. – May. – The electronic version of the printing publication. – URL: <http://csis.pace.edu/~ctappert/srd2008/c2.pdf> (access date: 22.08.2016).

11. Brocardo M.L. Authorship Verification for Short Messages using Stylometry [Electronic resource] / M.L. Brocardo, I. Traore, S. Saad, I. Woungang // University of Victoria. – Electronic data. – Victoria, BC, [s. a.]. – URL: [http://www.uvic.ca/engineering/ece/isot/assets/docs/Authorship\\_Verification\\_for\\_Short\\_Messages\\_using\\_Stylometry.pdf](http://www.uvic.ca/engineering/ece/isot/assets/docs/Authorship_Verification_for_Short_Messages_using_Stylometry.pdf) (access date: 22.08.2016).

12. Zheng R. A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques [Electronic resource] / R. Zheng, J. Li, H. Chen, Z. Huang // Journal of the American society for information science and technology. – 2006. – Vol. 57, № 3. – P. 378–393. – The electronic version of the printing publication. – URL:

[http://www.pages.drexel.edu/~jl622/docs/Journals/Zheng\\_2006JASIST\\_Authorship Identification.pdf](http://www.pages.drexel.edu/~jl622/docs/Journals/Zheng_2006JASIST_Authorship Identification.pdf) (access date: 22.08.2016).

13. Ragel R.G. Authorship detection of SMS messages using unigrams [Electronic resource] / R.G. Ragel, P. Herath, U. Senanayake // *Industrial and Information Systems (ICIIS) : 8th IEEE International Conference on 17–20 Dec. 2013. – 2013. – P. 387–392. – Doi: 10.1109/ICIInfS.2013.6732015.*

14. Павлов А.С. Исследование и разработка методов построения программных средств обнаружения текстового спама : автореф. дис. ... канд. физ.-мат. наук / А.С. Павлов. – М., 2012. – 15 с.

15. Павлов А.С. Исследование и разработка методов построения программных средств обнаружения текстового спама : дис. ... канд. физ.-мат. наук / А.С. Павлов. – М., 2011. – 133 с.

16. A reference collection for web spam / C. Castillo [et al.] // *ACM Sigir Forum 2006. – 2006. – Vol. 40, iss. 2. – P. 11–24.*

17. Gyöngyi Z.P. Applications of Web link analysis : PhD thesis / Z.P. Gyöngyi. – Stanford, 2008.

18. Gyöngyi Z.P. Link spam detection based on mass estimation / Z.P. Gyöngyi, P. Berkhin, H. Garcia-Molina, J. Pedersen // *Proceedings of the 32<sup>nd</sup> International Conference on Very Large Databases (VLDB). – 2006. – P. 439–450.*

19. Анисимов А.В. Методы вычисления мер семантической близости слов естественного языка / А.В. Анисимов, К.С. Лиман, А.А. Марченко // *Искусственный интеллект. – 2010. – № 3. – С. 170–175.*

20. Агеев М.С. Сложные задачи автоматической рубрикации текстов [Электронный ресурс] / М.С. Агеев, Б.В. Добров, Н.В. Лукашевич // *Научный сервис в сети Интернет : труды Всероссийской науч. конф. – Новороссийск, 2002. – Электрон. версия печатн. публ. – URL: [http://www.cir.ru/docs/ips/publications/2002\\_abrau\\_rubr.pdf](http://www.cir.ru/docs/ips/publications/2002_abrau_rubr.pdf) (дата обращения: 22.08.2015).*

21. Агеев М.С. Автоматическая рубрикация текстов: методы и проблемы // М.С. Агеев, Б.В. Добров, Н.В. Лукашевич // Ученые записки Казанского государственного университета. Сер. Физико-математические науки. – 2008. – Т. 150, кн. 4. – С. 25–40.

22. Дунаев Е.В. Автоматическая рубрикация web-страниц в интернет-каталоге с иерархической структурой / Е.В. Дунаев, А.А. Шелестов // Уральский федеральный университет имени первого Президента России Б.Н. Ельцина. – Электрон. дан. – [Екатеринбург, 2005]. – URL: [http://elar.urfu.ru/bitstream/10995/1419/1/ИМАТ\\_2005\\_20.pdf](http://elar.urfu.ru/bitstream/10995/1419/1/ИМАТ_2005_20.pdf) (дата обращения: 08.08.2016).

23. Фоменко А.Т. Методы статистического анализа исторических текстов (приложения к хронологии) : в 2 т. / А.Т. Фоменко. – М. : Крафт+Леан, 1999. – Т. 1. – 832 с.; Т. 2. – 908 с.

24. Кто написал «Тихий Дон»? : проблема авторства «Тихого Дона» / Г. Хьетсо [и др.]. – М. : Книга, 1989. – 186 с.

25. Анохин В. Жизнь под грифом «Секретно» [Электронный ресурс] / В. Анохин // Макспарк. Электрон. дан. – [Б. м.], 2015. – URL: <http://maxpark.com/community/5392/content/3241182> (дата обращения: 23.08.2016).

26. Хмелев Д. Краткая история разработки методик определения авторского стиля [Электронный ресурс] / Д. Хмелев // ЛингвоАнализатор. – Электрон. текст. дан. – [Б. м.], 1999. – URL: <http://www.rusf.ru/books/analysis/history.htm> (дата обращения: 23.08.2016).

27. Методы атрибуции [Электронный ресурс] // В поисках потерянного автора : этюды атрибуции. – Электрон. текст. дан. – [Б. м., б. г.]. – URL: [http://attribution.corneille-moliere.com/?p=attribution\\_methods&m=main&l=rus](http://attribution.corneille-moliere.com/?p=attribution_methods&m=main&l=rus) (дата обращения: 23.08.2016).

28. Литвинова Т.А. Установление характеристик (профилирование) автора письменного текста / Т.А. Литвинова // Филологические науки : вопросы теории и практики. – 2012. – № 2 (13). – С. 90–94.

29. Милов Л.В. От Нестора до Фонвизина: новые методы определения авторства / Л.В. Милов. – М. : Прогресс, 1994. – 443 с.

30. Родионова Е.С. Методы атрибуции художественных текстов / Е.С. Родионова // Структурная и прикладная лингвистика : межвуз. сб. / под ред. А.С. Герда. – СПб. : Изд-во С.-Петербур. ун-та, 2008. – Вып. 7. – С. 118–127.

31. Иванян Е.П. Общее языкознание. Теория языка [Электронный ресурс] : курс лекций / Е.П. Иванян. – 2-е изд., стер. – М. : Флинта, 2014. – Ч. 2. – 463 с. // Books.google.ru. – Электрон. дан. – [Б. м.], 2014. – URL: <https://books.google.ru/books?id=0jZ7AwAAQBAJ&printsec=frontcover&hl=ru#v=onepage&q&f=false> (дата обращения: 15.08.2016).

32. Батура Т.В. Формальные методы определения авторства текстов / Т.В. Батура // Вестник Новосиб. гос. ун-та. Сер. Информационные технологии. – 2012. – Т. 10, вып. 4. – С. 81–94.

33. Вязигина Н.В. Диагностика пола автора как задача автороведческой экспертизы [Электронный ресурс] / Н. В. Вязигина // Юрислингвистика: судебная лингвистическая экспертиза, лингвоконфликтология, юридико-лингвистическая герменевтика. Электрон. дан. – [Б. м.], 2012. – URL: 2012. [http://konference.siberia-expert.com/publ/konferencija\\_2012/doklad\\_s\\_obsuzhdeniem\\_na\\_sajte/diagnostika\\_pola\\_avtora\\_kak\\_zadacha\\_avtorovedcheskoj\\_ekspertizy/5-1-0-114](http://konference.siberia-expert.com/publ/konferencija_2012/doklad_s_obsuzhdeniem_na_sajte/diagnostika_pola_avtora_kak_zadacha_avtorovedcheskoj_ekspertizy/5-1-0-114) (дата обращения: 23.08.2016).

34. Резанова З.И. Задачи авторской атрибуции текста в аспекте гендерной принадлежности (к проблеме междисциплинарного взаимодействия лингвистики и информатики) / З.И. Резанова, А.С. Романов,

Р.В. Мещеряков // Вестник Томского государственного университета. – 2013. – № 370. – С. 24–28.

35. Родионова Е.С. Лингвистические методы атрибуции и датировки литературных произведений (к проблеме «Корнель–Мольер» : автореф. дис. ... канд. филол. наук / Е.С. Родионова. – СПб., 2008. – 24 с.

36. Напреенко Г.В. Идентификация текста по его авторской принадлежности на лексическом уровне (формально-количественная модель) / Г.В. Напреенко // Вестник Томского государственного университета. – 2014. – № 379. – С. 17–23.

37. Корпусные исследования письменной речи в решении задач судебного автороведения / Т.А. Литвинова, О.А. Литвинова, Е.В. Диброва, Е.С. Рыжкова // Филологические науки. Вопросы теории и практики. – 2015. – № 8-1 (50). – С. 107–113.

38. Шумская А.О. Идентифицирующие признаки текстовых сообщений при установлении автора / А.О. Шумская // Ползуновский вестник. – 2013. – № 2. – С. 265–266.

39. Красса С.И. Методика и инструментарий атрибуции текста в автороведческой экспертизе / С.И. Красса // Альманах современной науки и образования. – 2013. – № 10 (77). – С. 106–108.

40. Суровцова Т.Г. О построении статистических критериев для атрибуции авторства литературных текстов / Т.Г. Суровцова, С.П. Чистяков // Вестник Санкт-Петербургского университета. Сер. 10, Прикладная математика. Информатика. Процессы управления. – 2009. – Вып. 3. – С. 137–142.

41. Орлов Ю.Н. Методы статистического анализа литературных текстов / Ю.Н. Орлов, К.П. Осминин. – М. : Либроком, 2012. – 312 с.

42. Суровцова Т.Г. Многомерный количественный анализ и классификация текстов на основе линвостатистических характеристик :

автореф. дис. ... канд. техн. наук / Т.Г. Суровцова. – Петрозаводск, 2008. – 18 с.

43. Гудков В.Ю. N-граммы в лингвистике / В.Ю. Гудков, Е.Ф. Гудкова // Вестник Челябинского государственного университета. – 2011. – № 24 (239): Филология. Искусствоведение. – Вып. 57. – С. 69–71.

44. Мощенкова Д.С. Обзор программных продуктов разработанных для атрибуции художественных текстов [Электронный ресурс] / Д.С. Мощенкова, Д.А. Кривицкая, Н.С. Амосова // Молодежь и наука : сборник материалов X Юбилейной Всерос. науч.-техн. конф. студентов, аспирантов и молодых ученых с международным участием, посвященной 80-летию образования Красноярского края, 2014 г., Красноярск / Сибирский федеральный ун-т, 2014. – Электрон. такст. дан. – URL: [http://elib.sfu-kras.ru/bitstream/handle/2311/17293/s43\\_010.pdf?sequence=1&isAllowed=y](http://elib.sfu-kras.ru/bitstream/handle/2311/17293/s43_010.pdf?sequence=1&isAllowed=y) (дата обращения: 15.03.2016).

45. Малютов М.Б. Обзор методов и примеров атрибуции текстов / М.Б. Малютов // Обозрение прикладной и промышленной математики. – 2005. – Т. 12, № 1. – С. 41–77. – Электрон. версия печатн. публ. – URL: <http://people.oregonstate.edu/~kelberta/dima/misc/NewPapers/Malutov-marlowe-rus.pdf> (дата обращения: 23.08.2016).

46. Лингвоанализатор [Электронный ресурс]. – Электрон. дан. – [Б. м., б. г.]. – URL: [www.rusf.ru/books/analysis](http://www.rusf.ru/books/analysis) (дата обращения: 26.05.2016).

47. Текстология.ru [Электронный ресурс]. – Электрон. дан. – [Б. м., б. г.]. – URL: <http://www.textology.ru/> (дата обращения: 23.08.2016).

48. АОТ : автоматическая обработка текста [Электронный ресурс]. – Электрон. дан. – [Б. м., 2016]. – URL: <http://www.aot.ru/index.html> (дата обращения: 31.07.2016).

49. Луньков А.Д. Интеллектуальный анализ данных [Электронный ресурс] : учеб.-метод. пособие / А.Д. Луньков, А.В. Харламов // Саратовский

национальный исследовательский государственный университет имени Н.Г. Чернышевского. – Электрон. дан. – Саратов, [б. г.]. – URL: [http://elibrary.sgu.ru/uch\\_lit/1141.pdf](http://elibrary.sgu.ru/uch_lit/1141.pdf) (дата обращения: 13.08.2016).

50. Макаренко С.И. Интеллектуальные информационные системы : учеб. пособие / С.И. Макаренко. – Ставрополь : СФ МГГУ им. М.А. Шолохова, 2009. – 206 с.

51. Павлов С.Н. Системы искусственного интеллекта : учеб. пособие : в 2 ч. / С.Н. Павлов. – Томск : Эль Контент, 2011. – Ч. 1. – 176 с.

52. Романов А.С. Подходы к идентификации авторства текста на основе n-грамм и нейронных сетей [Электронный ресурс] / А.С. Романов // Современные информационные технологии : научная конференция молодых ученых ТПУ. – Электрон. текст. дан. – [Б.м.], 2011. – URL: <http://topetot.ru/191.html> (дата обращения: 18.08.2016).

53. Финн В.К. Об интеллектуальном анализе данных [Электронный ресурс] / В.К. Финн // Новости Искусственного интеллекта. – 2004. – № 3. – Электрон. версия печатн. публ. – URL: [www.raai.org/about/persons/finn/pages/finn\\_kdd.doc](http://www.raai.org/about/persons/finn/pages/finn_kdd.doc) (дата обращения: 08.08.2016).

54. Нейронные сети [Электронный ресурс] // Statsoft : электронный учебник по статистике. – Электрон. дан. – [Б. м., б. г.]. – URL: <http://www.statsoft.ru/home/textbook/modules/stneunet.html> (дата обращения: 08.08.2016).

55. Juola P. JGAAP: A System for Comparative Evaluation of Authorship Attribution [Electronic resource] / P. Juola // JDHCS 2009. – Vol. 1, № 1. – The electronic version of the printing publication. – URL: <https://letterpress.uchicago.edu/index.php/jdhcs/article/view/4/37> (access date: 22.08.2016).

56. Partyka J. More Advanced Stylometry with JGAAP and R-stylo [Electronic resource] / J. Partyka // Temple University. – Electronic data. –



Philadelphia, PA, [2015]. – URL: <http://sites.temple.edu/tudsc/2015/08/11/more-advanced-stylometry-with-jgaap-and-r-stylo/> (access date: 22.08.2016).

57. Развитие системы автоматического анализа текстов «СтилеАнализатор» / А.С. Красцова, В.В. Поддубный, О.Г. Шевелев, А.А. Фатыхов, О.В. Кукушкина, А.А. Поликарпов // Русский язык: исторические судьбы и современность : IV Международный конгресс исследователей русского языка, 20–23 марта 2010 г., Москва / Моск. гос. ун-т им. М.В. Ломоносова, филолог. фак. : труды и материалы. – М. : Изд-во Моск. ун-та, 2010. – С. 520–521.

58. Ирхин В. Фильтр Яндекса за неуникальный контент на сайте [Электронный ресурс] / В. Ирхин // Semantica. – Электрон. дан. – [М., б. г.]. – URL: <http://semantica.in/blog/filtr-yandeksa-za-neunikalnyj-kontent-na-sajte.html> (дата обращения: 08.08.2016).

59. Генерация текстов. Онлайн-сервисы и программы [Электронный ресурс] // Blog-Craft.ru. – Электрон. дан. – [Б. м., 2016]. – URL: <http://blog-craft.ru/generaciya-tekstov-onlajn-servisy-i-programmy/> (дата обращения: 08.08.2016).

60. Зеленков Ю.Г. Сравнительный анализ методов определения нечетких дубликатов для Web-документов [Электронный ресурс] / Ю.Г. Зеленков, И.В. Сегалович // Электронный библиотеки: перспективные методы и технологии, электронные коллекции : 9 всероссийская научная конференция, 15–18 октября 2007 г., Переславль-Залесский. – Электрон. дан. – [Б. м., 2007]. – URL: [http://rcdl2007.pereslavl.ru/papers/paper\\_65\\_v1.pdf](http://rcdl2007.pereslavl.ru/papers/paper_65_v1.pdf) (дата обращения: 01.06.2016).

61. Родненко В. Поиск нечетких дубликатов. Алгоритм шинглов для веб-документов [Электронный ресурс] / В. Родненко // Публикации. – Электрон. дан. – [Б. м., 2009]. – URL: <https://habrahabr.ru/post/65944/> (дата обращения: 08.08.2016).

62. Введение в общие цепи Маркова : учеб.-метод. пособие / А.В. Зорин [и др.]. – Н. Новгород : Нижегород. гос. ун-т, 2013. – 51 с

63. Автоматическое порождение текста [Электронный ресурс] // Tpl it : проект студентов специальности «Теоретическая и прикладная лингвистика» и магистрантов Иркутского государственного лингвистического университета. – Электрон. текст. дан. – [Иркутск], 2013. – URL: <http://tpl-it.wikispaces.com/Автоматическое+порождение+текста>. (дата обращения: 25.01.2013).

64. [ManHunter]. Генератор текста на основе цепей Маркова [Электронный ресурс] // Blog. just blog. – Электрон. дан. – [Б. м., 2010]. – URL: [http://www.manhunter.ru/webmaster/358\\_generator\\_teksta\\_na\\_osnove\\_sereyu\\_markova.html](http://www.manhunter.ru/webmaster/358_generator_teksta_na_osnove_sereyu_markova.html) (дата обращения: 08.08.2016).

65. Сравнение текстов на схожесть [Электронный ресурс]. – Электрон. дан. – [Б. м., б. г.]. – URL: <http://utext.rikuz.com/> (дата обращения: 08.08.2016).

66. Гасфилд Д. Строки, деревья и последовательности в алгоритмах: информатика и вычислительная биология / Д. Гасфилд ; пер. с англ. И.В. Романовского. – СПб. : Невский диалект; БХВ-Петербург, 2003. – 654 с.

67. Еще один журнал из «списка ВАК» опубликовал сгенерированную компьютером статью [Электронный ресурс] // Полит.РУ. – Электрон. текст. дан. – [Б. м.], 2004. – URL: <http://www.polit.ru/article/2009/04/01/erunda/> (дата обращения: 06.04.2013).

68. Лавренченко С.А. Порождающие грамматики [Электронный ресурс] / С.А. Лавренченко // Лавренченко точка ру. – Электрон. дан. – [Б. м., 2009]. – URL: <http://www.lavrencenko.ru/files/grammatics-11-lawrencenko.pdf> (дата обращения: 08.08.2016).

69. Что такое синонимайзер [Электронный ресурс] // Reraiteх.Ru. – Электрон. дан. – [Б. м., 2010–2013]. – URL: <http://www.reraiteх.ru/public/sininiмайzer.php> (дата обращения: 08.08.2016).

70. Что такое рерайтинг? Методы рерайтинга. Значение термина рерайтинг в Интернете [Электронный ресурс] // Аниматика. – Электрон. дан. – [Б. м., 2005–2013]. – URL: <http://animatika.ru/info/gloss/rewriting.html> (дата обращения: 08.08.2016).

71. Что такое копирайтинг? Виды копирайтинга. Значение термина копирайтинг [Электронный ресурс] // Аниматика. – Электрон. дан. – [Б. м., 2005–2013]. – URL: <http://animatika.ru/info/gloss/copywriting.html> (дата обращения: 08.08.2016).

72. [Александр] Генерация почти осмысленных текстов на Haskell [Электронный ресурс] // Записки программиста. – Электрон. дан. – [Б. м., 2011]. – URL: <http://eax.me/haskell-text-gen/> (дата обращения: 08.08.2016).

73. Козиев И. Синонимизатор / И. Козиев // Компьютерная грамматика русского языка: лексика, морфология, синтаксис. – Электрон. дан. – [Б. м., 2014]. – URL: [http://solarix.ru/for\\_developers/docs/synonymizer.shtml](http://solarix.ru/for_developers/docs/synonymizer.shtml) (дата обращения: 08.08.2016).

74. Web Spam: a Survey with Vision for the Archivist [Electronic resource] / A.A. Benezúr [et al.] // 8th International Web Archiving Workshop : conference proceedings. – Electronic data. – [Aarhus, 2008]. – URL: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.144.7624> (access date: 08.08.2016).

75. Detecting spam web pages through content analysis / A. Ntoulas [et al.] // Proceedings of the 15th international conference on World Wide Web – WWW'06. – Edinburgh, 2006. P. 83–92.

76. Назарова А. Google изнутри: борьба со спамом [Электронный ресурс] / А. Назарова // SEOnews. – Электрон. дан. – [М., 2013]. – URL: <https://www.seonews.ru/events/google-iznutri-borba-so-spamom/> (дата обращения: 08.08.2016).

77. How Search Works [Electronic resource]. – Electronic data. – [S. l., s. a.]. – URL: <http://www.google.com/insidesearch/howsearchworks/thestory/> (access date: 08.08.2016).

78. Петренко с. Как Google борется с поисковым спамом [Электронный ресурс] / С. Петренко // Searchengines.ru : энциклопедия интернет-маркетинга. – Электрон. дан. – [Б. м., 2011]. – URL: <https://www.searchengines.ru/009797.html> (дата обращения: 08.08.2016).

79. Некачественные сайты : Чем отличается качественный сайт от некачественного с точки зрения Яндекса? [Электронный ресурс] // Яндекс. – Электрон. дан. – [Б. м., б. г.]. – URL: <https://yandex.ru/support/webmaster/yandex-indexing/webmaster-advice.xml> (дата обращения: 08.08.2016).

80. Яндекс и поисковая оптимизация [Электронный ресурс] // Яндекс. – Электрон. дан. – [Б. м., 1997–2016]. – URL: <https://yandex.ru/company/rules/optimization/> (дата обращения: 08.08.2016).

81. Павлов А.С. Методы обнаружения поискового спама, порожденного с помощью цепей Маркова / А.С. Павлов, Б.В. Добров // Электронные библиотеки: перспективные методы и технологии, электронные коллекции : труды XI Всерос. науч. конф. – Петрозаводск, 2009. – Т. 1. – С. 311–317.

82. Павлов А.С. Метод обнаружения массово порожденных неестественных текстов на основе анализа тематической структуры / А.С. Павлов, Б.В. Добров // Вычислительные методы и программирование. – 2011. – Т. 12, вып. 3. – С. 58–72.

83. Поиск неестественных текстов / Е.А. Гречников [и др.] // ифровые библиотеки: продвинутые методы и технологии, цифровые коллекции : труды XI всерос. конф. – RCDL'2009. – Петрозаводск. 2009. – С. 306–308.

84. Зайцева А.А. Метод оценки качества текстов в задачах аналитического мониторинга информационных ресурсов / А.А. Зайцева,

С.В. Кулешов, С.Н. Михайлов // Труды СПИИРАН. – 2014. – Вып. 37. – С. 144–155.

85. Aharoni R., Koppel M., Goldberg Y. Automatic Detection of Machine Translated Text and Translation Quality / R. Aharoni, M. Koppel, Y. Goldberg // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. – Baltimore, Md, 2014. – Vol. 2. – P. 289–295.

86. Шумская А.О. Задачи идентификации искусственных текстов / А.О. Шумская // Научная сессия ТУСУР – 2013 : материалы Всероссийской научно-технической конференции студентов, аспирантов и молодых ученых, Томск, 16–18 мая 2013 г. : в 5 ч. – Томск : В-Спектр, 2012. – Ч. 4. – С. 224–226.

87. Романов А.С. Методика идентификации автора текста на основе аппарата опорных векторов / А.С. Романов // Доклады ТУСУРа. – 2009. – № 1 (19), ч. 2. – С. 36–42.

88. Романов А.С. Методика проверки однородности текста и выявления плагиата на основе метода опорных векторов и фильтра быстрой корреляции / А.С. Романов, З.И. Резанова, Р.В. Мещеряков // Доклады ТУСУРа. – 2014. – № 2 (32). – С. 264–269.

89. Романов А.С. Разработка и исследование математических моделей, методик и программных средств информационных процессов при идентификации автора текста : монография / А.С. Романов, А.А. Шелупанов, Р.В. Мещеряков. – Томск : В-Спектр, 2011. – 188 с.

90. Identifying the Authors of Suspect Email [Electronic resource] / M. Corney [et al.] // QUT ePrints. – Electronic data. – [S. l., s. a.]. – URL: <http://eprints.qut.edu.au/8021/1/CompSecurityPaper.pdf> (access date: 26.05.2016).

91. Исхакова А.О. Модель процесса формирования инвариантов классов текстов / А.О. Исхакова // Доклады ТУСУРа. – 2016. – № 3.

92. Романов А.С. Обобщенная методика идентификации автора неизвестного текста / А.С. Романов, А.А. Шелупанов, С.С. Бондарчук // Доклады ТУСУР. – 2010. – № 1 (21), ч. 1. – С. 108–112.

93. Шумская А.О. Об идентификации искусственно созданных текстов // А.О. Шумская, Р.В. Мещеряков / Шестая международная конференция по когнитивной науке : тезисы докладов. – Калининград, 2014. – С. 648–649.

94. Шумская А.О. Анализ текстовых признаков искусственных текстов, созданных на основе синонимизации / А.О. Шумская // Научная сессия ТУСУР – 2013 : материалы Всероссийской научно-технической конференции студентов, аспирантов и молодых ученых, Томск, 16–18 мая 2013 г. : в 5 ч. – Томск : В-Спектр, 2012. – Ч. 4. – С. 226–228.

95. Delirium 1.8 [Электронный ресурс] // Мониторинг Интернета. – Электрон. дан. – [М.], 2016. – URL: <http://ppc-seo.blogspot.ru/2007/10/delirium-18.html> (дата обращения: 01.04.2016).

96. Article Clone Easy – бесплатная программа для размножения статей [Электронный ресурс] // Энциклопедия поисковых систем. – Электрон. дан. – [Б.м.], 2013. – URL: <http://forum.searchengines.ru/showthread.php?t=317633> (дата обращения: 17.04.2016).

97. Валгина Н.С. Теория текста / Н.С. Валгина. – М. : Логос, 2003. – 191 с.

98. Николина Н.А. Филологический анализ текста : учеб. пособие / Н.А. Николина. – М. : Академия, 2003. – 256 с.

99. Шумская А.О. Выбор параметров для идентификации искусственно созданных текстов / А.О. Шумская // Доклады ТУСУРа. – 2013. – №2 (28). – С. 126–128.

100. Шумская А.О. Определение искусственных текстов на основе поиска часто употребляемых слов и устойчивых словосочетаний /

А.О. Шумская // Седьмая международная конференция по когнитивной науке : тезисы докладов. – Светлогорск, 2016. – С. 647–648.

101. Национальный корпус русского языка [Электронный ресурс]. Электрон. дан. – М., 2003–2016. – URL: <http://www.ruscorpora.ru/index.html> (дата обращения: 09.12.2015).

102. Загоруйко Н.Г. Прикладные методы анализа данных и знаний / Н.Г. Загоруйко. – Новосибирск : ИМ СО РАН, 1999. – 270 с.

103. Гмурман В.Е. Теория вероятностей и математическая статистика / В.Е. Гмурман. – М. : Высшее образование, 2008. – 479 с

104. Исхакова А.О. Метод определения искусственных текстов на основе расчета меры принадлежности к инвариантам / А.О. Исхакова // Труды СПИИРАН. – 2016.

105. Шумская А.О. Оценка эффективности метрик расстояния Евклида и расстояния Махаланобиса в задачах идентификации происхождения текста / А.О. Шумская // Доклады ТУСУРа. –2013. – № 3 (29). – С. 141–145.

106. Shumskaya A.O. Using Euclidean and Mahalanobis distances while solving the problem of the text origin identification, scientific paper / A.O. Shumskaya // Interactive systems: Problems of Human – Computer Interaction : collection of scientific papers. – Ulyanovsk : USTU, 2015. – P. 211–217.

107. Shumskaya (Iskhakova) A.O. / The effectiveness of the Euclidean and Mahalanobis distances while solving the problem of the text origin identification / A.O. Shumskaya // Modern informatization problems : Proceedings of the XIX International Open Science Conference (Yelm, WA, Usa, January 2014) / ed. O.Ja. Kravets. – Yelm, WA : Science Book Publishing House, 2014. – P. 73–78.

108. SQL инъекция в MySQL сервере [Электронный ресурс] // Securitylab. – Электрон. дан. – [Б. м.], 2004. – URL: <http://www.securitylab.ru/contest/212083.php> (дата обращения: 15.02.2016).

109. Юникод, UTF-8 – современная кодировка [Электронный ресурс] // Veb.name. – Электрон. дан. – [Б. м.], 2011. – URL: <http://veb.name/index.php?document=yunikodUTF8sov1> (дата обращения: 15.02.2016).

110. Функции для работы с Многобайтными строками [Электронный ресурс] // The PHP Group. – Электрон. дан. – [Б. м.], 2016. – URL: <http://php.net/manual/ru/ref.mbstring.php> (дата обращения: 15.08.2016).

111. Прокопюк А. Разработка → PHP RUtils — небольшая библиотека для обработки русского текста [Электронный ресурс] / А. Прокопюк // Хабрахабр. – Электрон. дан. – [Б. м.], 2013. – URL: <https://habrahabr.ru/post/198544/> (дата обращения: 23.08.2016).

112. RussianSentimentAnalyzer API [Electronic resource] / J. Partyka // SemanticAnalyzer. – Electronic data. – [S. l., s. a.]. – URL: <http://semanticanalyzer.info/blog/russiansentimentanalyzer-api/> (access date: 22.08.2016).

113. Наши технологии: определение темы и ключевых терминов текста [Электронный ресурс] // Семантическое зеркало / Ашманов и партнеры. – Электрон. дан. – [М., б. г.]. – URL: <http://sm.ashmanov.com/product/> (дата обращения: 30.08.2016).

114. Математические функции [Электронный ресурс] // PHP.RU – Сообщество PHP-программистов. – Электрон. дан. – [Б. м.], 2016. – URL: <https://php.ru/manual/ref.math.html> (дата обращения: 08.08.2016).

115. Вычисления с увеличенной точностью (GNU Multiple Precision) [Электронный ресурс] // The PHP Group. – Электрон. дан. – [Б. м.], 2016. – URL: <http://php.net/manual/ru/book.gmp.php> (дата обращения: 19.08.2016).



116. Иваненко С. Введение в SNMP [Электронный ресурс] / С. Иваненко // Network.xsp.ru. – Электрон. дан. – [Б. м.], 2010. – URL: [http://network.xsp.ru/6\\_1.php](http://network.xsp.ru/6_1.php) (дата обращения: 30.08.2016).

117. Томсон Л. Разработка Web-приложений на PHP и MySQL : пер. с англ. / Л. Томсон, Л. Веллинг. – 2-е изд., испр. – СПб. : ДиаСофтЮП, 2003. – 672 с.

118. Федосеева А. Спецификация языка Perl [Электронный ресурс] / А. Федосеева // Citforum. – Электрон. дан. – [Б. м., б. г.]. – URL: <http://citforum.ru/database/cnit/p2.shtml> (дата обращения: 20.08.2016).

119. Стальная CMS // Information Security / Информационная безопасность. – 2008 – № 5. – С. 34.

120. Модуль ngx\_http\_upstream\_module [Электронный ресурс] // Nginx. – Электрон. дан. – [Б. м., б. г.]. – URL: [http://nginx.org/ru/docs/http/ngx\\_http\\_upstream\\_module.html](http://nginx.org/ru/docs/http/ngx_http_upstream_module.html) (дата обращения: 24.08.2016).

121. CSS: настройки CSS, стилизация HTML-элементов с использованием расширенных классов, и передовая система разметки [Электронный ресурс]. – Электрон. дан. – [Б. м., б. г.]. – URL: <http://bootstrap-3.ru/css.php> (дата обращения: 29.08.2016).

## Приложение А

## Свидетельство о государственной регистрации программы для ЭВМ

РОССИЙСКАЯ ФЕДЕРАЦИЯ



**СВИДЕТЕЛЬСТВО**  
о государственной регистрации программы для ЭВМ  
**№ 2015663136**

**Auth\_stat. Программное обеспечение для расчета  
статистических значений текстовых характеристик**

Правообладатель: **Федеральное государственное бюджетное  
образовательное учреждение высшего профессионального  
образования «Томский государственный университет систем  
управления и радиозлектроники» (ТУСУР) (RU)**

Авторы: **Шумская Анастасия Олеговна (RU),  
Мещераков Роман Валерьевич (RU)**

Заявка № **2015660397**  
Дата поступления **29 октября 2015 г.**  
Дата государственной регистрации  
в Реестре программ для ЭВМ **11 декабря 2015 г.**

Руководитель Федеральной службы  
по интеллектуальной собственности



 **Г.П. Ивлиев**

## Приложение Б

Значения характеристик для текстовых выборок разного происхождения

Текстовая характеристика	Естественный текст	Искусственный текст (синонимизация)	Искусственный текст (метод цепей Маркова)
1. Средняя длина слов	5,83	6,213	5,851
2. Среднее количество знаков пунктуации на 1000 символов	31,742	29,035	21,545
3. Частота 100 популярных биграмм букв на 1000 символов	201,269	112,562	184,36
4. Частота служебных слов на 1000 символов	34,691	25,702	31,865
5. Частота неопределенных местоимений на 1000 символов	1,054	0,473	0,711
6. Частота коротких слов (менее 4 символов) на 1000 символов	52,974	40,613	51,056
7. Частота длинных слов (более 7 символов) на 1000 символов	51,743	51,766	50,168
8. Количество уникальных слов на 1000 символов	64,804	101,659	66,155
9. Среднее число слов в предложении;	9,113	9,987	16,549
10. Количество грамматических ошибок на 1000 символов	0,01	6,215	1,002
11. Количество предложений в тексте на 1000 символов	109,812	100,2	60,356
12. Количество сложноподчиненных предложений на 1000 символов	68,655	62,082	31,658
13. Доля сложноподчиненных предложений;	62,525	61,962	52,45
14. Количество вопросительных предложений на 1000 символов	1,414	1,358	0,787
15. Количество восклицательных предложений на 1000 символов	0,113	0,102	0,005
16. Доля вопросительных и восклицательных предложений	0,013	0,014	0,013
17. Частота 100 популярных слов на 1000 символов	49,001	32,882	37,104
18. Частота 100 популярных 2-грамм слов на 1000 символов	9,1	3,554	4,196
19. Частота 100 популярных 3-грамм слов на 1000 символов	3,612	0,365	1,015
20. Количество слов в семантическом ядре	66,025	95,645	75,268
21. Наличие единства тематики в разных частях текста	2,862	1,135	1,578

## Приложение В

Графическое представление значений характеристик для текстовых выборок  
разного происхождения

На гистограмме для обозначения нормированных в интервале  $[0;1]$  значений измеренных характеристик текстов используются следующая нумерация:

1. средняя длина слов;
2. среднее количество знаков пунктуации на 1000 символов;
3. частота 100 популярных биграмм букв на 1000 символов;
4. частота служебных слов на 1000 символов;
5. частота неопределенных местоимений на 1000 символов;
6. частота коротких слов (менее 4 символов) на 1000 символов;
7. частота длинных слов (более 7 символов) на 1000 символов;
8. количество уникальных слов на 1000 символов;
9. среднее число слов в предложении;
10. количество грамматических ошибок на 1000 символов;
11. количество предложений в тексте на 1000 символов;
12. количество сложноподчиненных предложений на 1000 символов;
13. доля сложноподчиненных предложений;
14. количество вопросительных предложений на 1000 символов;
15. количество восклицательных предложений на 1000 символов;
16. доля вопросительных и восклицательных предложений;
17. частота 100 популярных слов на 1000 символов;
18. частота 100 популярных 2-грамм слов на 1000 символов;
19. частота 100 популярных 3-грамм слов на 1000 символов;
20. количество слов в семантическом ядре;
21. наличие единства тематики в разных частях текста.



## Приложение Г

## Результаты расчетов для оценки различительной способности

Результаты расчетов приведены в таблице настоящего приложения. Используются следующие обозначения:

$M_1(X_k)$  – математическое ожидание величины значения характеристики  $X_k \in X$  для выборки естественных текстов;

$M_2(X_k)$  – математическое ожидание величины значения характеристики  $X_k \in X$  для выборки искусственных текстов, созданных с помощью синонимизации;

$M_3(X_k)$  – математическое ожидание величины значения характеристики  $X_k \in X$  для выборки искусственных текстов, созданных с помощью метода, основанного на цепях Маркова;

$s_1(X_k)$  – среднеквадратичное отклонение величины значения характеристики  $X_k \in X$  для выборки естественных текстов;

$s_2(X_k)$  – среднеквадратичное отклонение величины значения характеристики  $X_k \in X$  для выборки искусственных текстов, созданных с помощью синонимизации;

$s_3(X_k)$  – среднеквадратичное отклонение величины значения характеристики  $X_k \in X$  для выборки искусственных текстов, созданных с помощью метода, основанного на цепях Маркова.

$k$	$X_k$	$M_1(X_k)$	$M_2(X_k)$	$M_3(X_k)$	$ M_1(X_k) - M_2(X_k) $	$ M_1(X_k) - M_3(X_k) $	$s_1(X_k) + s_2(X_k)$	$s_1(X_k) + s_3(X_k)$
1	Средняя длина слов	5,83	6,213	5,851	0,383	0,021	0,512	0,453
2	Среднее количество знаков пунктуации на 1000 символов	31,742	29,035	21,545	2,707	10,197	1,911	4,125
3	Частота 100 популярных биграмм букв на 1000 символов	201,269	112,562	184,36	88,707	16,909	11,105	10,546
4	Частота служебных слов на 1000 символов	34,691	25,702	31,865	8,989	2,826	4,603	2,542
5	Частота неопределенных местоимений на 1000 символов	1,054	0,473	0,711	0,581	0,343	0,997	0,254
6	Частота коротких слов (менее 4 символов) на 1000 символов	52,974	40,613	51,056	12,361	1,918	3,598	1,878
7	Частота длинных слов (более 7 символов) на 1000 символов	51,743	51,766	50,168	0,023	1,575	6,228	5,443
8	Количество уникальных слов на 1000 символов	64,804	101,659	66,155	36,855	1,351	9,845	0
9	Среднее число слов в предложении;	9,113	9,987	16,549	0,874	7,436	0,213	1,647
10	Количество грамматических ошибок на 1000 символов	0,01	6,215	1,002	6,205	0,992	1,258	0,846
11	Количество предложений в тексте на 1000 символов	109,812	100,2	60,356	9,612	49,456	4,017	2,184
12	Количество сложноподчиненных предложений на 1000 символов	68,655	62,082	31,658	6,573	36,997	4,711	7,115
13	Доля сложноподчиненных предложений;	62,525	61,962	52,45	0,563	10,075	4,569	5,015
14	Количество вопросительных предложений на 1000 символов	1,414	1,358	0,787	0,056	0,627	0,051	0,245
15	Количество восклицательных предложений на 1000 символов	0,113	0,102	0,005	0,011	0,108	0,010	0,009
16	Доля вопросительных и восклицательных предложений	0,013	0,014	0,013	0,001	0	0,315	0,403
17	Частота 100 популярных слов на 1000 символов	49,001	32,882	37,104	16,119	11,897	4,326	4,012
18	Частота 100 популярных 2-грамм слов на 1000 символов	9,1	3,554	4,196	5,546	4,904	3,524	2,548
19	Частота 100 популярных 3-грамм слов на 1000 символов	3,612	0,365	1,015	3,247	2,597	1,956	1,254
20	Количество слов в семантическом ядре	66,025	95,645	75,268	29,62	9,243	7,841	6,874
21	Наличие единства тематики в разных частях текста	2,862	1,135	1,578	1,727	1,284	0,984	0,845

## Приложение Д

## Акты внедрения результатов диссертационного исследования

Экз. № 1

**Общество с ограниченной ответственностью  
«Агентство медиарешений»**

ИНН/КПП 7017340248/701701001  
тел. +7-3822-783-720, +7-903-915-0016

АКТ от «25» июля 2016 г.  
о внедрении результатов диссертационной работы  
на соискание ученой степени кандидата технических наук  
Исхаковой Анастасии Олеговны

Общество с ограниченной ответственностью «Агентство медиарешений» занимается сбором и агрегированием информации новостного характера. Одним из направлений деятельности компании является модерирование контента для новостных интернет-порталов, в том числе обработка текстовой информации на предмет выявления сообщений, сгенерированных автоматически. Результаты диссертационной работы Исхаковой А.О. используются в ООО «Агентство медиарешений» для автоматизации процедур анализа текстов и последующей фильтрации данных, подлежащих публикации.

Разработанное Исхаковой А.О. программное обеспечение (ПО) «TextOrigin» позволяет в полной мере использовать предложенную ей методику и инварианты естественных и искусственных текстов для определения текстов, сгенерированных автоматически. Отличительной особенностью ПО является возможность автоматической фильтрации входных данных при достижении заданного порога, характеризующего текст как искусственный.

Реализация программного средства на базе веб-ориентированной технологии позволила создать кроссплатформенное решение, которое может быть интегрировано в действующие веб-ресурсы, обладающие потребностью анализа входных данных по критерию искусственности их происхождения. Возможность применения «TextOrigin» в качестве фильтра для модулей приема онлайн-обращений или модулей публикации записей в социальных сетях, обуславливает высокую степень практической значимости разработанного ПО.

В результате эксплуатации программного средства в период с февраля по май 2016 года было проанализировано 1712 текстов. Средний объем текста составил 2400 знаков. Установлено, что уровень ошибок первого рода составил 2,9%, а уровень ошибок второго рода 1,2%. При этом удалось не только значительно снизить трудозатраты, но и существенно повысить качество обработки текстового материала за счет снижения зависимости от человеческого фактора.

Настоящий акт составлен в 3 (трех) экземплярах.

Директор ООО «АМР»



Е.Ю. Суходолина



«УТВЕРЖДАЮ»  
 Генеральный директор  
 М.В.Федюкин  
 « 2 августа 2016 г.



А К Т № 110 от 2.08.2016г.

о внедрении результатов диссертационной работы  
 на соискание ученой степени кандидата технических наук  
 Исаковой Анастасии Олеговны

Комиссия в составе:

председатель - заместитель генерального директора Мельников С.Ю.,  
 члены комиссии - начальник отдела разработки программного обеспечения и электронного обучения Пичкур К.Б., руководитель общей группы Летунова М.Е.

составили настоящий акт о том, что результаты диссертационной работы Исаковой Анастасии Олеговны использованы в производственной деятельности Общества с ограниченной ответственностью «Лингвистические и информационные технологии» в виде математических моделей, методов и алгоритмов выявления искусственно сгенерированных текстов и программной реализации этих алгоритмов.

Использование указанных результатов позволяет повысить качество автоматизированного анализа исследуемых материалов на ряде мировых языков и языков народов России.

Результаты внедрялись при выполнении работ по договору N3/11.16 от 20.02.2016 с ФГУП «Ростовский НИИ радиосвязи».

Общество с ограниченной ответственностью «Лингвистические и информационные технологии» занимается научными исследованиями и разработкой программного обеспечения в области прикладной и компьютерной лингвистики. Развитие методов обработки неструктурированной текстовой информации является одним из важных направлений деятельности Общества, в частности, интерес представляет анализ текстов, полученных из различных интернет-источников.

Председатель комиссии

 С.Ю. Мельников

члены комиссии

 К.Б. Пичкур  
 М.Е. Летунова

**ТУСУР**

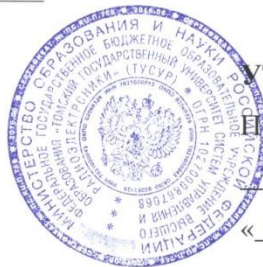
Министерство образования и науки Российской Федерации  
Федеральное государственное бюджетное  
образовательное учреждение высшего образования  
«ТОМСКИЙ ГОСУДАРСТВЕННЫЙ  
УНИВЕРСИТЕТ СИСТЕМ УПРАВЛЕНИЯ  
И РАДИОЭЛЕКТРОНИКИ»

ОКПО 02069326, ОГРН 1027000867068,  
ИНН 7021000043, КПП 701701001

тел: (382 2) 510-530  
факс: (382 2) 513-262, 526-365  
e-mail: [office@tusur.ru](mailto:office@tusur.ru)  
http:// [www.tusur.ru](http://www.tusur.ru)

пр. Ленина, 40, г. Томск, 634050

№ \_\_\_\_\_

**УТВЕРЖДАЮ**

Проректор ТУСУР по учебной работе

П.Е. Троян

« 26 09 2016г.

**АКТ**

**о внедрении результатов диссертационной работы  
Исхаковой Анастасии Олеговны в учебный процесс**

Комиссией в составе:

Председатель комиссии – Давыдова Е.М, к.т.н., декан факультета безопасности ТУСУР

Члены комиссии:

Костюченко Е.Ю., к.т.н., доцент каф. КИБЭВС ТУСУР;

Евсютин О.О., к.т.н., доцент каф. БИС ТУСУР;

составлен настоящий акт о нижеследующем.

Результаты диссертационной работы А.О. Исхаковой «Методика и программное средство определения искусственно созданных текстов» используются в учебном процессе на факультете безопасности ТУСУР при чтении лекций и в практических занятиях по дисциплинам «Дискретная математика» и «Математическая статистика и теория вероятностей» для подготовки специалистов по защите информации, обучающихся по специальностям «10.05.03 – Информационная безопасность автоматизированных систем» и «10.05.04 – Информационно-аналитические системы безопасности».

В курсе по дисциплине «Дискретная математика» используются результаты диссертационной работы Исхаковой А.О. по формальному описанию задачи определения происхождения текста и моделирования

процесса создания инвариантов, а также разделы, посвященные созданию искусственных текстов с использованием аппарата формальных грамматик и Марковских цепей.


Результаты исследования процесса создания инвариантов для решения задач обработки текстовых данных, в частности модель данного процесса, предложенная Исхаковой А.О., используются в практических занятиях по дисциплине «Математическая статистика и теория вероятностей» для формирования навыков применения статистического аппарата в решении прикладных задач.

Кроме того, в ходе выполнения групповых проектов и научно-исследовательских работ студенты факультета безопасности имеют возможность ознакомиться с результатами проведенных исследований искусственных текстов, а также использовать разработанное программное обеспечение.

Рассмотрение студентами результатов диссертационной работы Исхаковой А.О. в рамках практических занятий по указанным дисциплинам позволяет сформировать навыки применения изучаемых аппаратов представления математических структур и проведения статистических расчетов для решения прикладной задачи анализа данных.

Настоящий акт составлен в 3 (трех) экземплярах.


к.т.н., декан факультета  
безопасности  
Давыдова Е.М.

  
«29» 08 2016 г.

к.т.н., доцент каф. КИБЭВС  
Костюченко Е.Ю.

  
«29» 08 2016 г.

к.т.н., доцент каф. БИС  
Евсютин О.О.

  
«29» 08 2016 г.