

Санкт–Петербургский государственный университет

**Дубовик Анна Романовна**

**Автоматическое определение стилистической  
принадлежности текстов по их статистическим параметрам**

**Магистерская диссертация**

Направление «Лингвистика»,  
Образовательная программа  
«Прикладная и экспериментальная  
лингвистика»,  
профиль «Компьютерная лингвистика  
и интеллектуальные технологии»

Научный руководитель:  
доц., к.ф.н. Митрофанова О.А.

Санкт–Петербург

2017

## СОДЕРЖАНИЕ

|  |    |
|--|----|
| ВВЕДЕНИЕ .....   | 3  |
| ГЛАВА 1. Теоретические основания для автоматической стилистической<br>диагностики текстов на русском языке.....              | 6  |
| 1.1. Функциональные стили русского языка и их характеристики .....   | 6  |
| 1.1.1. Научный стиль .....   | 13 |
| 1.1.2. Художественный стиль.....   | 14 |
| 1.1.3. Деловой стиль .....   | 14 |
| 1.1.4. Публицистический стиль.....   | 16 |
| 1.2. Классификация текстовой информации.....   | 17 |
| 1.2.1. Методы дискриминантного анализа .....   | 21 |
| 1.2.2. Оценка качества работы алгоритма классификации текстовой<br>информации .....  | 24 |
| ГЛАВА 2. КОМПЬЮТЕРНЫЙ ИНСТРУМЕНТ ДЛЯ ПРОВЕДЕНИЯ<br>СТАТИСТИЧЕСКОЙ ОБРАБОТКИ РУССКОЯЗЫЧНЫХ ТЕКСТОВ .....                      | 26 |
| 2.1. Используемое программное обеспечение.....   | 26 |
| 2.2. Требования к входным данным.....  | 26 |
| 2.3. Алгоритм работы компьютерного инструмента статистической<br>обработки текстов .....                                     | 28 |
| 2.4. Интерфейс компьютерного инструмента проведения статистической<br>обработки текстов .....                                | 31 |
| ГЛАВА 3. ЭКСПЕРИМЕНТАЛЬНАЯ ПРОВЕРКА ВОЗМОЖНОСТИ<br>АВТОМАТИЧЕСКОЙ СТИЛИСТИЧЕСКОЙ КЛАССИФИКАЦИИ<br>РУССКОЯЗЫЧНЫХ ТЕКСТОВ..... | 34 |
| 3.1. Подготовка корпусов.....  | 34 |
| 3.2. Подбор характеризующих признаков.....   | 34 |
| 3.3. Ход экспериментов.....  | 36 |
| 3.4. Анализ данных .....   | 37 |
| 3.4.1. Анализ лексико–морфологических индексов .....   | 37 |
| 3.4.2. Анализ материала на основе данных о частеречной сочетаемости  | 41 |
| 3.4.3. Параметры длины слова и длины предложения .....   | 45 |

|   |    |
|---|----|
| 3.5. Инструмент автоматического определения стилистической принадлежности текстов .....                     | 47 |
| 3.5.1. Описание алгоритма стилистической принадлежности текстов...  | 47 |
| 3.5.2. Оценка качества работы модуля автоматического определения стилистической принадлежности текстов..... | 50 |
| ЗАКЛЮЧЕНИЕ .....  | 53 |
| СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ .....  | 55 |
| ПРИЛОЖЕНИЕ А. Код программы автоматического определения стилистической принадлежности текстов .....         | 63 |
| ПРИЛОЖЕНИЕ Б. Перечень текстов, использованных при создании корпусов .....                                  | 69 |

## ВВЕДЕНИЕ

В последние годы очень быстрыми темпами развивается область обработки естественных языков (англ. *Natural Language Processing, NLP*). Во многом это связано с тем, что с каждым годом объём текстовой информации, используемой человечеством, увеличивается, и растёт потребность в более эффективных алгоритмах обработки и анализа документов, написанных на естественных языках. Особо важную роль играет возможность классифицировать получаемую информацию, используя компьютерные инструменты.

Таким образом, **актуальность** выбранной темы обусловлена необходимостью разработки новых методов автоматической обработки текстов и востребованностью новых методов классификации текстовой информации при помощи компьютеров.

**Материалом** исследования послужили данные четырёх корпусов текстов, представляющих различные функциональные стили русского языка (художественный, научный, деловой и публицистический). Объём каждого корпуса – 500 тыс. словоупотреблений.

**Целью** настоящего исследования является разработка компьютерного инструмента автоматического определения стилистической принадлежности текстов.

В соответствии с поставленной целью работы сформулированы следующие **задачи** исследования:

- проанализировать ряд теоретических вопросов, связанных с выделением функциональных стилей текстов современного русского языка;
- выделить отличительные количественные характеристики для каждого стиля;
- сравнить собранные корпуса текстов, опираясь на выделенные характеристики;

- проанализировать различные методы классификации текстовой информации;
- создать авторский компьютерный инструмент определения стилистической принадлежности для текстов на русском языке;
- использовать разработанный компьютерный инструмент для обработки сформированных корпусов;
- оценить возможность автоматического определения стилистической принадлежности текстов по выделенным характеристикам.

**Методы исследования**, использованные в работе, включают стилистический и статистический анализ корпусов текстов русского языка.

**Программное обеспечение**, необходимое для проведения исследования, было подготовлено на языке python версии 2.7.13 и представляет собой реализацию алгоритма статистической обработки текста и определения его стилистической принадлежности. Всем словоупотреблениям в обрабатываемых текстах приписываются грамматические характеристики. Затем данные проходят автоматическую обработку, и проводится их исследовательская интерпретация. Систематизация полученных данных приводит к выявлению статистических параметров текста, а также к выявлению параметров, являющихся характеризующими для текстов, принадлежащих к разным функциональным стилям. Затем проводится определение стилистической принадлежности текста при помощи модуля стилистической диагностики, использующего выявленные ключевые параметры.

**Теоретическая значимость** исследования определяется тем, что в результате анализа корпуса текстов нам удаётся получить ценные данные о статистических характеристиках текстов различных функциональных стилей.

**Практическая значимость** работы заключается в том, что её результаты могут быть использованы в информационно–поисковых системах и при обработке больших объёмов текстовых данных.

**Апробация исследования:** основные положения исследования и полученные экспериментальные данные были представлены в докладе на XIX Открытой конференции студентов-филологов 20 апреля 2016 года.

**Структура квалификационной работы:** работа состоит из введения, трёх глав, заключения, списка использованной литературы и приложений.

## **ГЛАВА 1. Теоретические основания для автоматической стилистической диагностики текстов на русском языке.**

В проведённом нами исследовании решается задача автоматической стилистической диагностики текстов на русском языке. Предлагаемое решение основывается, с одной стороны, на лингвистических параметрах, определяющих стилистическую принадлежность текстов, а с другой – на алгоритмах автоматической классификации текстовой информации.

### **1.1. Функциональные стили русского языка и их характеристики**

Теория стилей русского языка восходит к учению о трёх штилях М.В. Ломоносова, описанному им в сочинении «Предисловие о пользе книг церковных в российском языке». В основу этого учения Ломоносовым был положен экспрессивно–жанровый принцип: стили литературного языка соотносились с жанрами художественной прозы, поэзии и драмы. Соотношение было следующим [Ломоносов, 1952, с. 589–590]:

- 1) высокий штиль – оды, трагедии, героические поэмы, гимны, ораторские речи (произведения, в которых описываются возвышенные чувства или исторические события);
- 2) средний штиль – драмы, элегии, эклоги, стихи, сатиры, письма, научные сочинения (произведения, просвещающие читателя, рассказывающие о современных ему событиях, о жизни известных людей);
- 3) низкий штиль – комедии, песни, эпиграммы, басни, дружеские письма и записки (произведения, предназначенные для развлечения).

Каждому из этих штилей соответствовали определённые лексические нормы: например, в произведениях низкого штиля широко употреблялись простонародные слова и жаргонизмы, в то время как использование подобных лексических средств в высоком штиле было недопустимо. В

произведениях среднего штиля можно было встретить обычные разговорные слова, но запрещалось использование бранных и уничижительных слов (за исключением тех случаев, когда того требовало действие).

А.Х. Востоков в работе «Русская грамматика» рассуждает о трёх типах речи (имея в виду общенародный язык). Согласно предложенной им классификации, речь бывает [Востоков, 1831, с. 4]:

- 1) важная (книжный язык);
- 2) обыкновенная (разговорный язык);
- 3) простонародная (просторечие).

Следует обратить внимание, что в основу классификации М.В. Ломоносова были положены разновидности литературного языка, а в основу теории А.Х. Востокова – разновидности общенародного языка.

Становление современного учения о функциональных стилях русского языка связано с трудами ряда учёных: Г.О. Винокура [Винокур, 1959], В.В. Виноградова [Виноградов, 1980], М.Н. Кожиной [Кожина, 1984], Б.Н. Головина [Головин, 1988] и др.. Ещё в первой четверти XX века, рассуждая о культуре языка, Г.О. Винокур писал: «речь устная и письменная, ораторская и разговорная, канцелярская и парламентская, докладная записка или указ, беседа с приятелем и дипломатический обмен любезностями, язык в прозе и стихах – все эти языковые задания ... требуют и своих средств выполнения» [Винокур, 1929, с. 27]. Г.О. Винокур заложил основы современной стилистики, в центре которой – учение о функционально–стилевой дифференциации литературного языка.

К настоящему времени это учение получило хорошее развитие, однако ряд вопросов остаётся дискуссионным. Наибольшие затруднения вызывает, как ни странно, центральный вопрос стилистики – выделение различных стилей языка. А.И. Горшков в учебном пособии «Русская стилистика» указывает на то, что функциональные стили в языке могут быть выделены на основании функций языка, в соответствии со сферами функционирования



языка и с опорой на структуру текста [Горшков, 2006, с. 265]. Рассмотрим эти принципы подробнее.

Первый подход – выделение стилей языка с опорой на его функции. Например, В.В. Виноградов в своей работе «Стилистика. Теория поэтической речи. Поэтика» писал: «При выделении таких важнейших общественных функций языка, как общение, сообщение и воздействие, могли бы быть в общем плане структуры языка разграничены такие стили: обиходно–бытовой стиль (функция общения); обиходно–деловой, официально–документальный и научный (функция сообщения); публицистический и художественно–беллетристический (функция воздействия)» [Виноградов, 1963, с. 6]. Д.Э. Розенталь в «Практической стилистике русского языка» также подчёркивает, что «язык как явление социальное выполняет различные функции <...> Для реализации этих функций исторически сложились и оформились отдельные разновидности языка» [Справочник по русскому языку. Практическая стилистика, 2001, с. 12]. Стоит, однако, отметить, что функции языка трудно чётко отделить друг от друга, а функция общения присутствует во всех случаях его употребления.

Во–вторых, стили можно выделять в соответствии со сферами функционирования языка. Этим принципом обычно руководствуются авторы научной и учебной литературы по стилистике; в частности, именно он описывается в популярных учебниках для высших учебных заведений: «Стилистике русского языка» М.Н. Кожиной [Кожина, 1993] и «Стилистике русского языка» под редакцией Н.М. Шанского [Стилистика русского языка, 1989]. Согласно этому принципу, каждой сфере деятельности человека соответствует определённый стиль. «Научную сферу обслуживает научный стиль, деловую – официально–деловой, обиходно–разговорную – разговорный, а в сфере массовой информации используется публицистический стиль» [Там же, с. 146]. В «Новом словаре методических терминов и понятий» также даётся следующее определение функциональных стилей: «стили, выделяемые в соответствии с основными функциями языка,

связанными с той или иной сферой деятельности человека» [Азимов, Щукин, 2009, с. 342–343]. Этот принцип выделения стилей русского языка получил гораздо большее распространение, но и он не лишён недостатков. Во-первых, сферы деятельности людей весьма многообразны, причём в рамках одной области может выделяться несколько дополнительных, более конкретных областей. Также следует помнить, что сферы деятельности человека постоянно претерпевают изменения: например, с появлением Интернета изменилась структура СМИ, а с ними – и сама журналистика [Калмыков, 2005]. Таким образом, традиционно выделяемых стилей языка гораздо меньше, чем выделяемых областей деятельности, и для того, чтобы подобрать для каждой сферы деятельности (и каждой конкретной области) соответствующий стиль, приходится дополнительно выделять подстили, жанры, поджанры и т.п. Во-вторых, при выделении характеристик стиля возможна некоторая их «подгонка» под внеязыковые факторы: например, известно, что существует сфера дипломатической деятельности человека, значит, существует и соответствующий ей стиль языка – официально-деловой (или, например, его дипломатический подстиль). Так как существует деловой стиль, можно выделить и описать его специфические признаки; также можно определить специфические характеристики и для его дипломатического подстиля.

Третий принцип выделения стилей – с опорой на структуру текста. Например, В.В. Одинцов в «Стилистике текста» пишет: «в структуре текстов обнаруживаются два основных композиционно-стилистических типа изложения, подачи содержания: научно-деловой, информационно-логический, с одной стороны, и беллетризованный, экспрессивный – с другой» [Одинцов, 1980, с. 78]. В современной литературе этот принцип не получил достаточного распространения, а некоторые исследователи полагают, что он является обобщением первых двух подходов, так как структура текста во многом определяется тем, на что текст нацелен и в какой сфере он будет функционировать [Мурашова, 2017, с. 86]. В

«Лингвистическом энциклопедическом словаре» под редакцией В.Н. Ярцевой также читаем: «функциональный стиль – это разновидность литературного языка, в которой язык выступает в той или иной социально значимой сфере общественно–речевой практики людей и особенности которой обусловлены особенностями общения в данной сфере. Наличие функционального стиля связывают также с различием функций, выполняемых языком» [Лингвистический энциклопедический словарь, 1990].

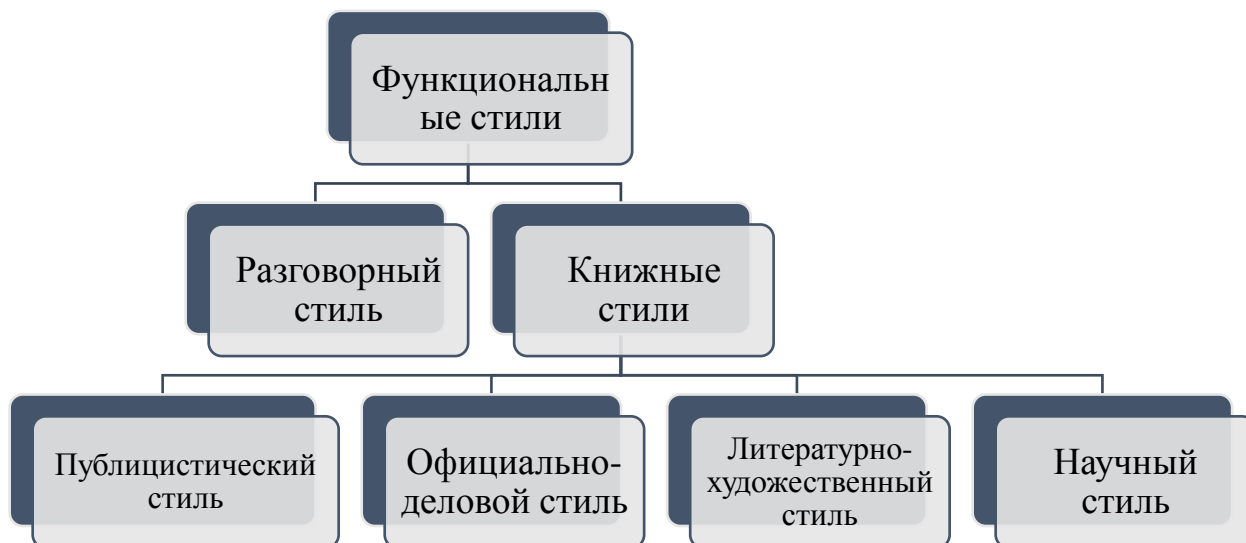
Как мы видим, два принципа выделения стилей из трёх, описанных выше, содержат явные недостатки, а оставшийся не проработан. Из этого следует существующая и по нынешний день неопределённость при описании системы стилей современного русского языка.

На данный момент не существует однозначного ответа на вопрос о том, какие функциональные стили есть в современном русском языке. Разные исследователи выделяют различное их число.

М.Н. Кожина утверждает, что обычно выделяют пять основных функциональных стилей: научный, официально–деловой, публицистический, художественный и разговорно–бытовой [Кожина, 1993, с. 160].

В «Стилистике русского языка» под редакцией Н.М. Шанского рассматриваются научный, официально–деловой, публицистический и разговорный стили и особо – язык художественной литературы, который не включается в систему функциональных стилей [Стилистика русского языка, 1989, с. 155–208].

Д.Э. Розенталь представляет систему стилей в виде следующей схемы:



*Рис. 1. Функциональные стили русского языка по Д.Э. Розенталю.*

К схеме даётся примечание: «Литературно–художественный (художественно–беллетристический) стиль принадлежит к числу книжных стилей, и его место в левой части схемы, но в связи с присущим ему своеобразием он не попадает в один ряд с другими книжными стилями» [Справочник по русскому языку. Практическая стилистика, 2001, с. 381].

Стоит отметить, что во всех приведённых системах выделяются научный, официально–деловой и публицистический стиль. Эти стили фиксируются практически во всех работах по стилистике.

Следует также обратить особое внимание на то, что М.Н. Кожина выделяет «художественный стиль», Н.М. Шанский говорит о языке художественной литературы, а в схеме Д.Э. Розенталя присутствует «литературно–художественный стиль», но подчеркивается, что «он не попадает в один ряд с другими книжными стилями». Это связано с другим вопросом стилистики: положением языка художественной литературы в системе разновидностей употребления русского языка. Обычно в литературе по стилистике рассматриваются следующие спорные вопросы:

1) Можно ли считать язык художественной литературы одним из функциональных стилей, или надо говорить о нем как об особой разновидности употребления языка;

2) Какова связь между языком художественной литературы и литературным языком и какое понятие шире – «язык художественной литературы» или «литературный язык».

Более подробно с мнениями различных исследователей можно ознакомиться, например, в работах В.В. Виноградова [Виноградов 1955], Р.А. Будагова [Будагов, 1967], Ф.М. Березина, Б.Н. Головина [Березин, Головин, 1979] и Д.Н. Шмелёва [Шмелёв, 1977]. Мы в своей работе рассматриваем язык художественной литературы как одну из разновидностей литературного языка наряду с научным, официально–деловым и публицистическим функциональными стилями.

В работах М.Н. Кожиной и Н.М. Шанского в одном ряду с научным, официально–деловым и публицистическим стилями рассматривается и «разговорный стиль». Д.Э. Розенталь рассматривает этот стиль особо, отделяя его от «книжных стилей», но всё же говорит не о разговорном языке, соотношенном с литературным языком, а именно о «разговорном стиле» [Справочник по русскому языку. Практическая стилистика, 2001, с. 12]. В самом деле, как замечает А.И. Горшков, в составе литературного языка не может быть выделено разговорной разновидности, так как литературное и разговорное употребление языка противопоставлены друг другу [Горшков, 2006, с. 269]. В своей работе мы не рассматриваем «разговорный стиль» как стиль литературного русского языка.

Экстралингвистические факторы, условия, в которых используется язык, влияют на отбор речевых средств, вероятность употребления тех или иных лексем, грамматических форм и конструкций [Функциональные стили и формы речи, 1993, с. 3].

В нашей работе рассматриваются тексты четырёх стилей: научного, официально–делового, художественного и публицистического. Рассмотрим кратко характеристики каждого из них.

### **1.1.1. Научный стиль**

Сфера общественной деятельности, в которой функционирует научный стиль – наука, причём преимущественно используемая форма речи – письменная. Основная функция данного функционального стиля – сообщение, фиксация результатов познания мира. Специфическая черта текстов этого стиля – понятийная точность, подчеркнутая логичность. Основные жанры произведений, использующих научный стиль: научная монография, научная статья, научно–учебная проза (учебники, учебные и методические пособия и т.п.), научно–технические произведения (инструкции, правила техники безопасности и т.д.), аннотации, рефераты, научные доклады, лекции, научные дискуссии.

Научный стиль имеет ряд особенностей, проявляющихся независимо от характера наук (естественных, точных, гуманитарных) и жанровых различий (монография, научная статья, доклад, учебник и т.д.), что дает возможность говорить о специфике стиля в целом:

- 1) терминологичность, господство обобщённо–отвлечённой лексики;
- 2) специфические параметры распределения частей речи (наименьшая среди всех стилей частотность предикативных форм глагола; общее превалирование имён; преобладание глаголов в форме настоящего времени);
- 3) специфические параметры распределения синтаксических структур (частотность сложных предложений, причастных оборотов, безличных предложений; преобладающее использование комбинированных словосочетаний) [Функциональные стили и формы речи, 1993, с. 37–44].

### **1.1.2. Художественный стиль**

Сфера использования художественного стиля – художественная литература. Основная функция текстов, принадлежащих к данному стилю – воздействие через индивидуально–образное моделирование мира. Специфическая черта текстов этого стиля – эстетическая значимость всех языковых элементов и образность речи. Для художественного стиля характерны:

- 1) лексическое богатство;
- 2) употребление преимущественно семантически конкретных существительных;
- 3) обилие глаголов говорения, разнообразных частиц и местоимений–существительных;
- 4) высокая частотность форм именительного и родительного падежей имён существительных (см. таблицы в [Функциональные стили и формы речи, 1993,, с. 103]);
- 5) преимущественное использование простых словосочетаний.

### **1.1.3. Деловой стиль**

Деловой стиль обслуживает административно–правовую сферу деятельности, соотносится с познавательно–регулирующей и предписывающей деятельностью сознания [Теплова, 2011, с. 665]. Этот стиль служит для оформления документов: законов, приказов, постановлений и др.. Сфера использования официально–делового стиля – право.

Очевидно, что наиболее распространённая форма существования этого стиля – письменная. Основными требованиями, предъявляемыми к тексту официально–делового стиля, являются точность (недопущение двусмысленности), стандартизированность (строгая композиция текста, точная форма подачи фактов), отсутствие эмоциональной оценки

сообщаемой информации [Функциональные стили и формы речи, 1993, с. 12].

Для этого стиля характерны [Там же, с. 12–13]:

- 1) стандартное расположение материала;
- 2) широкое использование терминологии (деловой, юридической), а также официальной, канцелярской лексики и фразеологии, включение в текст сложносокращенных слов, аббревиатур;
- 3) частое употребление отглагольных существительных, производных предлогов (*на основании, в отношении, в соответствии с, в целях, за счет и др.*), производных союзов (*вследствие того что, ввиду того что, в связи с тем что, в силу того что и др.*), а также различных устойчивых словосочетаний, служащих для связи частей сложного предложения (*на случай, если ...; на том основании, что ...; по той причине, что ...; с тем условием, что ...; таким образом, что ...; то обстоятельство, что ...; тот факт, что ... и т. п.*);
- 4) использование номинативных предложений с перечислением;
- 5) использование цепочек родительных падежей (см. пример в работе [Теплова, 2011, с. 665]: *компетенция органов государственной власти субъекта Российской Федерации в области жилищных отношений: установление порядка определения размера дохода и стоимости имущества, находящегося в собственности членов семьи и подлежащего налогообложению в целях признания граждан малоимущими и предоставления им жилых помещений муниципального жилищного фонда*);
- 6) тенденция к употреблению сложноподчинённых предложений, отражающих логическое подчинение одних фактов другим;
- 7) почти полное отсутствие эмоционально–экспрессивных речевых средств.



#### 1.1.4. Публицистический стиль

Публицистический стиль обслуживает сферу политико–идеологических социальных отношений, соотносится с познавательно–оценивающей работой сознания [Функциональные стили и формы речи, 1993, с. 65]. Этот стиль употребляется в сферах политико–идеологических, общественных и культурных отношений. Его цель – привлечь внимание адресата. Поэтому средства достижения выразительности и экспрессивности в публицистическом стиле имеют большее значение, чем в других стилях [Бикмуканова, 2014, с. 36–37].

Основные жанры произведений, относящихся к этому стилю – статья, очерк, репортаж, интервью и др.. Стоит особо отметить, что наравне с письменной формой данного стиля получила широкое распространение и его устная форма – это стиль радиопередач, выступлений по телевидению, а также выступлений на митингах, собраниях и т.п. [Функциональные стили и формы речи, 1993, с. 65].

Для этого стиля характерны:

- 1) наличие общественно–политической лексики и фразеологии, переосмысление лексики других стилей (в частности, терминологической) для целей публицистики;
- 2) использование характерных для данного стиля клише. К клише относятся различные стёртые метафоры, речевые штампы и устойчивые эпитеты (*по данным социологического опроса, жест доброй воли, мрачные прогнозы, горячая поддержка* и пр.) [Бондарь, 2016, с. 53];
- 3) использование изобразительно–выразительных средств языка, в частности средств стилистического синтаксиса (риторические вопросы и восклицания, параллелизм, повторы, инверсия) [Современная газетная публицистика. Проблемы стиля, 1987, с. 34];

Публицистический стиль представляет собой область межстилевых взаимодействий. Он сочетает в себе характеристики других функциональных стилей языка (в особенности научного и художественного), что приводит к усложнению его структуры [Москвин, 2006, с. 622]. Это делает задачу правильной автоматической классификации текстов, принадлежащих к нему, более трудоёмкой.

Как мы видим, лексико–грамматические и синтаксические характеристики функциональных стилей довольно различны. Следовательно, на основании этих характеристик можно отличать тексты одного стиля от текстов, относящихся к другому стилю. Мы предполагаем, что возможно подобрать такие комбинации параметров, которые позволят однозначно определять стиль исследуемого текста. Выявление подобных комбинаций позволит провести автоматическую классификацию русскоязычных текстов по разным стилям речи.

## **1.2. Классификация текстовой информации**

Одна из задач нашей работы – построение компьютерного инструмента автоматической стилистической диагностики текстов на русском языке, опирающегося, с одной стороны, на лингвистические параметры, определяющие стилистическую принадлежность текстов, а с другой – на алгоритмы автоматической классификации текстовой информации.

Классификация документов – это одна из задач информационного поиска, заключающаяся в автоматическом отнесении документа к одной или нескольким категориям на основании его содержания [Вартан, 2015, с. 31]. В англоязычной литературе употребляется термин *text categorization* (букв. *категоризация текстов*; см., например, работы [Cleuziou, 2007; Sebastiani, 2005; Apte, Damerau, Weiss, 1994]).

В наше время задача автоматической классификации документов приобретает всё большую актуальность: объём обрабатываемой текстовой

информации постоянно увеличивается, а использование классификаторов позволяет значительно ускорить процесс её анализа и сортировки.

А.С. Епрев в работе «Автоматическая классификация текстовых документов» указывает, что в настоящее время задача классификации имеет практическое применение в следующих областях [Епрев, 2010, с. 65]:

- 1) фильтрация спама;
- 2) составление тематических каталогов;
- 3) контекстная реклама;
- 4) сужение области поиска в поисковых системах;
- 5) системы документооборота;
- 6) автоматический перевод текстов (снятие омонимии).

А.Г. Васнецов определяет задачу классификации следующим образом. Имеется некоторое множество объектов, разделенное на непересекающиеся классы. Для части объектов данного множества (для некоторого конечного подмножества объектов данного множества) известно, к каким классам они принадлежат. Это подмножество называется выборкой. Для остальных объектов данного множества принадлежность к определённому классу не определена. Задача заключается в построении алгоритма, способного классифицировать произвольный объект из исходного множества [Васнецов, 2015].

Существует три основных подхода к задаче классификации текстов [Manning, Raghavan, Schütze, 2008, с. 255].

### **1. Ручная классификация.**

Из общей характеристики данного типа классификации следует, что он осуществляется вручную, без применения компьютера. Основной недостаток ручной классификации – невозможность её применения в случаях, когда необходимо классифицировать большое количество документов с высокой скоростью.

## **2. Классификация с использованием правил.**

При использовании данного вида классификации человеку (например, специалисту, знакомому с описываемой предметной областью и обладающему навыком написания регулярных выражений), необходимо составить ряд правил, по которым можно отнести текст к той или иной категории. Например, одно из таких правил может выглядеть следующим образом: "если текст содержит слова *производная* и *уравнение*, то его следует отнести к категории *математика*". Эти правила затем автоматически применяются к поступающим документам для их классификации.

При использовании правил процесс классификации автоматизируется и, следовательно, количество обрабатываемых документов практически не ограничено. Более того, построение правил вручную может дать лучшую точность классификации, чем при машинном обучении [Гулин, 2013, с. 115–121]. Однако создание и поддержание правил в актуальном состоянии требует постоянных усилий специалиста.

## **3. Классификация с использованием машинного обучения**

В этом подходе набор правил (или критерий принятия решения) текстового классификатора вычисляется автоматически из обучающих данных (другими словами, производится обучение классификатора). Обучающие данные – это некоторое количество хороших образцов документов из каждого класса. В машинном обучении сохраняется необходимость ручного приписывания класса документу. Но разметка является более простой задачей, чем написание правил. Кроме того, разметка может быть произведена в обычном режиме использования системы. Таким образом, классификация текстов, основанная на машинном обучении, является примером обучения с учителем, где в роли учителя выступает человек, задающий набор классов и размечающий обучающее множество.

В математической статистике задачи классификации называются также задачами дискриминантного анализа. Аппарат дискриминантного анализа разрабатывался с середины XX в. многими учеными (см., к примеру, работы Р. Фишера [Fisher, 1936], У.Р. Клекки [Klecka, 1980], Г.Д. Гарсона [Garson, 2012] и др.).

В основе метода дискриминантного анализа лежит понятие дискриминантной модели, которая оптимально разделяет множество объектов на подмножества и проводит классификацию новых объектов в тех случаях, когда неизвестно заранее, к какому из существующих классов они принадлежат [Панова, Денисова, 2014, с. 33].

В общем виде алгоритм дискриминантного анализа выглядит следующим образом [Большаков, Каримов, 2007, с.293–294]:

1. Определение возможного набора дискриминантных функций (англ. *discriminant functions*) или комбинаций независимых дискриминантных переменных.
2. Экспериментальная проверка существования между группами значимых различий с точки зрения выделенных функций или переменных.
3. Определение набора переменных, которые являются наиболее характеризующими для каждой группы.
4. Отнесение объектов к одной из заданных групп (собственно классификация), опираясь на значения переменных.
5. Оценка качества классификации (с использованием метрик точности, полноты, F–меры, FPR и др.).

Успешный опыт применения процедур этого вида анализа на текстовом материале описан в целом ряде работ (см., например, работы [Андреев, 2002; Ермолаева, 2009; Bagavandas, Manimannan, 2008]).

Рассмотрим основные методы дискриминантного анализа.

### **1.2.1. Методы дискриминантного анализа**

К основным методам дискриминантного анализа относятся линейный дискриминант Фишера, канонический дискриминантный анализ (он же линейный дискриминантный анализ), логистическая регрессия и деревья принятия решений. Рассмотрим подробнее каждый из них.

#### **1.2.1.1. Линейный дискриминант Фишера**

Линейный дискриминант Фишера был впервые предложен Р.Фишером в 1936 году в работе «The Use of Multiple Measurements in Taxonomic Problems» [Fisher, 1936].

Приведём описание данного метода, предложенное К.В. Воронцовым в курсе лекций по машинному обучению.

Предположим, что ковариационные матрицы классов одинаковы и равны  $\Sigma$ . Такой случай соответствует наилучшему разделению классов по дискриминанту Фишера (в первоначальном значении). Тогда статистический подход приводит к линейному дискриминанту, и именно этот алгоритм классификации в настоящее время часто понимается под термином линейный дискриминант Фишера.

Простота классификации линейным дискриминантом Фишера – одно из главных достоинств алгоритма: в случае с двумя классами в двумерном признаковом пространстве разделяющей поверхностью будет прямая. Если классов больше двух, то разделяющая поверхность будет кусочно–линейной. Но главным преимуществом алгоритма является уменьшение эффекта плохой обусловленности ковариационной матрицы при недостаточных данных [Воронцов, 2010, с. 30].

### **1.2.1.2. Канонический дискриминантный анализ**

При использовании канонического дискриминантного анализа классификационные корни и функции определяются каноническими корреляциями. Они задают новое пространство, в котором определяется центроид каждой группы, а объекты (точки) относятся к той группе элементов, к центроиду которой они находятся ближе всего. Максимальное число функций будет равно числу совокупностей минус один или числу переменных в анализе в зависимости от того, какое из этих чисел меньше [StatSoft. Электронный учебник по статистике].

Отбор признаков для проведения канонического дискриминантного анализа может осуществляться несколькими способами:

- 1) методом пошаговой регрессии – на каждом шаге проведения данной процедуры происходит либо включение признака в модель, либо его исключение;
- 2) оценкой уровня связи между исследуемым признаком и остальными признаками, уже включёнными в модель;
- 3) вычислением частного значения лямбды Уилкса – показателя степени важности признака для проведения классификации (насколько ухудшится результат классификации, если исключить этот признак из модели);
- 4) определением вероятности нулевой гипотезы, предполагающей, что при удалении признака точность классификации не изменится.

Этот вариант анализа допускает простую и очевидную графическую интерпретацию, что является большим преимуществом канонического дискриминантного анализа перед линейным.

### **1.2.1.3. Логистическая регрессия**

Логистическая регрессия или логит–регрессия (англ. logit model) — это статистическая модель, используемая для предсказания вероятности

возникновения некоторого события по значениям множества признаков. При этом используют несколько предсказывающих переменных, которые могут быть или числовыми, или категориальными [Ng, 2011, с. 16].

При решении задач классификации считается, что объект  $x$  можно отнести к классу  $y=1$ , если предсказанная моделью вероятность  $P\{y=1|x\} > 0,5$ , и к классу  $y=0$  в противном случае. Получающиеся при этом правила классификации являются линейными классификаторами.

#### **1.2.1.4. Дерево решений**

Дерево принятия решений (дерево классификации или регрессионное дерево) — средство поддержки принятия решений, использующееся в статистике и анализе данных для прогнозных моделей. Структура дерева представляет собой «листья» и «ветки». На «ветках» дерева решения записаны атрибуты, от которых зависит целевая функция, в «листьях» записаны значения целевой функции. Каждый лист представляет собой возможный выход алгоритма при обработке некоторых входных данных размера  $n$ .

Работу алгоритма с конкретными входными данными размером  $n$  можно представить как проход по дереву от «корня» до «листа»; количество сравнений при выполнении алгоритма равно количеству пройденных рёбер [Левитин, 2006, с. 409–417].

Подобные деревья решений широко используются в интеллектуальном анализе данных. Цель состоит в том, чтобы создать модель, которая предсказывает значение целевой переменной на основе нескольких переменных на входе.

Т.В. Зайцева отмечает следующие преимущества метода деревьев принятия решений [Зайцева и др., 2013, с. 122]:

1. Метод интуитивно понятен и прост в интерпретации;
2. Деревья решений дают возможность извлекать правила из базы данных на естественном языке;



3. Метод надёжен, хорошо работает даже в том случае, если были нарушены первоначальные предположения, включенные в модель.

4. Метод позволяет работать с большим объёмом информации без проведения специальных подготовительных процедур.

Для построения инструмента автоматического определения стилистической принадлежности текстов нами было принято решение использовать метод деревьев принятия решений.

### **1.2.2. Оценка качества работы алгоритма классификации текстовой информации**

Для оценки работы классификаторов чаще всего используются метрики точности (precision) и полноты (recall). Иногда они используются как самостоятельные метрики, иногда – как основа для производных метрик (например, F-меры).

При оценке качества классификатора необходимо учитывать не только те случаи, в которых он правильно определил принадлежность объекта к классу, но и случаи, когда он совершил ошибку. Введём следующие обозначения:

1) **TP** (true positive) — истинно–положительное решение (документ относится к классу A, классификатор отнёс его к классу A);

2) **FP** (false positive) — ложно–положительное решение (документ не относится к классу A, классификатор отнёс его к классу A);

3) **TN** (true negative) — истинно–отрицательное решение (документ не относится к классу A, классификатор не отнёс его к классу A);

4) **FN** (false negative) — ложно–отрицательное решение (документ относится к классу A, классификатор не отнёс его к классу A).

**Точность системы** в пределах класса – это доля документов, действительно принадлежащих данному классу, относительно всех документов, которые система отнесла к этому классу:

$$Precision = \frac{TP}{(TP + FP)}$$

**Полнота системы** – это доля найденных классификатором документов, принадлежащих классу, относительно всех документов этого класса в тестовой выборке:

$$Recall = \frac{TP}{(TP + FN)}$$

**F-мера (F-measure)** – характеристика, которая позволяет дать оценку одновременно по точности и полноте:

$$F - measure = \frac{1}{\alpha \frac{1}{Precision} + (1 - \alpha) \frac{1}{Recall}}, \alpha \in [0,1],$$

где  $\alpha$  – коэффициент, задающий соотношение весов точности и полноты. Когда  $\alpha = 0.5$ , F-мера придаёт одинаковый вес обеим характеристикам.

## ГЛАВА 2. КОМПЬЮТЕРНЫЙ ИНСТРУМЕНТ ДЛЯ ПРОВЕДЕНИЯ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ РУССКОЯЗЫЧНЫХ ТЕКСТОВ

### 2.1. Используемое программное обеспечение

В рамках диссертационного исследования нами был создан компьютерный инструмент для статистической обработки текстов на русском языке. Разработанная нами программа написана на языке программирования python 2.7.1 – наличие NLTK (пакета библиотек и программ для символьной и статистической обработки естественного языка) облегчает создание программ по обработке текстов на естественном языке.

Для корректной работы программы также необходимо воспользоваться морфологическим анализатором NLTK4RUSSIAN (находится в открытом доступе на сайте <https://github.com/named-entity/nltk4russian>). Подробнее о принципе работы данного программного средства см. работу [Паничева и др., 2015].

### 2.2. Требования к входным данным

Как уже указывалось выше, текст, который предполагается проанализировать при помощи нашей программы, необходимо предварительно обработать гибридным морфологическим анализатором NLTK4RUSSIAN, интегрирующим теггеры NLTK и морфологический анализатор PyMorphy2. На вход программе подаётся файл на русском языке с расширением .xml, сохранённый в кодировке UTF-8, или путь к папке, содержащей подобные файлы.

Рассмотрим реализацию морфологической и лексико-грамматической разметки, осуществляемой морфоанализатором, на примере слова *сел* в контексте *Он сел в низкое кресло у стола и вытянул свои длинные ноги*:

Он 1 сел 1 d VERB,perf,intr,masc,sing,past,indc в низкое кресло у  
стола и вытянул свои длинные ноги.

В аннотации отражены порядковый номер слова в предложении (*1*, так как нумерация слов в предложении начинается с 0), разделитель (*d*) словоформа (*сел*) и грамматические теги *VERB,perf,intr,masc,sing,past,indc* (глагол, совершенный вид, непереходный, мужской род, единственное число, прошедшее время, изъявительное наклонение).

Строки, содержащие знак препинания, обозначаются тегом *PNCT*; начало и конец каждого предложения обозначаются тегами *sent* и */sent* соответственно.

Примеры строк файла, обработанного морфоанализатором:

- Глагол: 1 спросил      1 d VERB,perf,tran,masc,sing,past,indc
- Междометие: 0 ой      0 d INTJ
- Местоимение–существительное:  
0 он      0 d NPRO,masc,3per,Anph,sing,nomn
- Наречие: 5 по–китайски      5 d ADVB
- Предлог: 2 в      2 d PREP
- Прилагательное: 7 опытный      7 d ADJF,Qual,masc,sing,nomn
- Союз: 3 и      3 d CONJ
- Существительное: 9 стола 9 d NOUN,inan,masc,sing,gent
  - Существительное – имя:  
4 валя 3 d NOUN,anim,femn,Ms–f,Name,sing,nomn
  - Существительное – фамилия:  
4 петров 4 d NOUN,anim,masc,Sgtm,Surn,sing,nomn
- Частица: 0 не      0 d PRCL
- Числительное: 2 два      2 d NUMR,masc,nomn
- Числительное–прилагательное:  
3 первую      3 d ADJF,Anum,femn,sing,accs.

Все примеры взяты из целевых корпусов, использованных в исследовании.

Результат использования морфоанализатора представляет собой файл, в котором каждая строка – либо тег начала/конца предложения, либо пунктуационный знак, либо словоформа с аннотацией (см. примеры выше).

### **2.3. Алгоритм работы компьютерного инструмента статистической обработки текстов**

В разработанном нами модуле реализуется алгоритм подсчёта статистических параметров указанного пользователем текста с расширением .xml.

С полным кодом разработанной программы можно ознакомиться в Приложении А.

При запуске программы пользователю необходимо передать приложению несколько параметров. Первый параметр – путь к файлу с расширением .xml, который необходимо обработать. Также пользователь может указать путь к папке, в которой содержатся интересующие его файлы – в этом случае программой будут последовательно обработаны все файлы с расширением .xml, содержащиеся в указанной папке. При указании несуществующего пути к файлу или папке программа выведет в консоль сообщение *«Пожалуйста, введите путь к файлу / папке с файлами»* и не начнёт работу.

Второй параметр – путь к файлу, в который требуется записать статистику по обработанным файлам. Если указанного пользователем файла не существует, он будет создан. Если пользователь не передаст программе никакой аргумент, результаты обработки файлов будут записаны в документ *«D:/Result.xml»*.

Третий параметр – запись полной статистики по всем обработанным текстам в файл. Если пользователь указывает на необходимость записи статистики, программа обрабатывает необходимые переменные и записывает их в конец файла с результатами обработки документов. В противном случае общая статистика не сохраняется.

После того, как пользователь ввёл необходимые параметры и нажал «Enter», приложение начинает обработку указанных файлов. Обработка каждого файла включает его открытие, построчное прочтение и вычисление статистических параметров содержащегося в нём текста.

В модуле нашей программы, производящем статистическую оценку текста, подсчитываются следующие параметры:

- 1) общее число слов в тексте и количество букв в них;
- 2) среднее число слов в предложении;
- 3) число слов, относящихся к разным частям речи, и их доли относительно общего числа слов в обрабатываемом документе;
- 4) количество конструкций, в которых друг за другом следуют два слова с морфологической характеристикой «существительное» (в частности, количество таких конструкций, в которых второе слово находится в форме родительного падежа), а также количество случаев, когда друг за другом следуют слова с характеристиками «глагол» и «существительное»;
- 5) соотношение динамичности и статичности в тексте (подробнее об этом параметре см. п. 3.4.2. и работу [Антонова, Клышинский, Ягунова, 2011]).

Полученные данные записываются в соответствующие переменные.

Затем происходит запись полученных данных в файл, указанный пользователем (или в файл, заданный по умолчанию). Для каждого обработанного документа программа сохраняет следующие параметры:

- 1) Среднее число слов в предложении (а также минимальное и максимальное значение данного индекса);
- 2) Среднее число букв в одном слове (а также минимальное и максимальное значение данного индекса);
- 3) Число групп «существительное + существительное» (в т.ч. число групп «существительное + существительное в родительном падеже») и групп «глагол + существительное»;

- 4) Значение параметра соотношения динамичности и статичности для данного текста;
- 5) Общее число слов в тексте;
- 6) Число существительных, глаголов, прилагательных, личных местоимений и частиц в тексте (а также их долю относительно общего числа слов в тексте);

Затем, если пользователь указал, что нет необходимости записывать в файл статистику по всем обработанным документам, программа завершает работу. В противном случае происходит запись следующих параметров:

- 1) Средняя доля существительных (по всем файлам);
- 2) Средняя доля глаголов (по всем файлам);
- 3) Средняя доля прилагательных (по всем файлам);
- 4) Средняя доля частиц (по всем файлам);
- 5) Средняя доля личных местоимений (по всем файлам);
- 6) Среднее число конструкций «глагол + существительное» (по всем файлам);
- 7) Среднее число конструкций «существительное + существительное» (по всем файлам);
- 8) Среднее число конструкций «существительное + существительное в родительном падеже» (по всем файлам);
- 9) Среднее значение параметра соотношения динамичности и статичности (по всем файлам);
- 10) Средняя длина предложений;
- 11) Средняя длина слов.

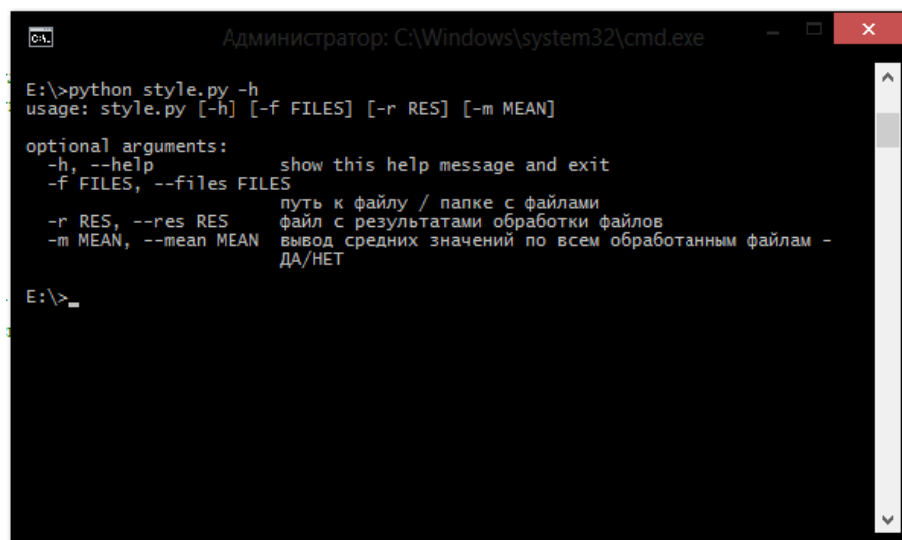
Для каждого параметра также указывается его минимальное и максимальное значения.

После записи данных индексов в файл программа выводит в консоль сообщение о том, что процесс обработки файлов завершён, и прекращает работу.

## 2.4. Интерфейс компьютерного инструмента проведения статистической обработки текстов

Интерфейс — совокупность возможностей, способов и методов взаимодействия двух систем (любых, а не обязательно являющиеся вычислительными или информационными), устройств или программ для обмена информацией между ними, определённая их характеристиками, характеристиками соединения, сигналов обмена и т. п. В случае, если одна из взаимодействующих систем — человек, чаще говорят лишь о второй системе, то есть об интерфейсе той системы, с которой человек взаимодействует [18, с. 223–224].

Созданная нами программа представляет собой консольное приложение с текстовым интерфейсом командной строки. Для того, чтобы запустить утилиту, пользователь должен вызвать интерпретатор Python, указать название файла с кодом программы, передать необходимые параметры и нажать «Enter». Пример запуска утилиты с параметром *-h* (вызов справки) показан на рисунке 2.



```
Администратор: C:\Windows\system32\cmd.exe
E:\>python style.py -h
usage: style.py [-h] [-f FILES] [-r RES] [-m MEAN]

optional arguments:
  -h, --help            show this help message and exit
  -f FILES, --files FILES  путь к файлу / папке с файлами
  -r RES, --res RES       файл с результатами обработки файлов
  -m MEAN, --mean MEAN   вывод средних значений по всем обработанным файлам -
                          ДА/НЕТ

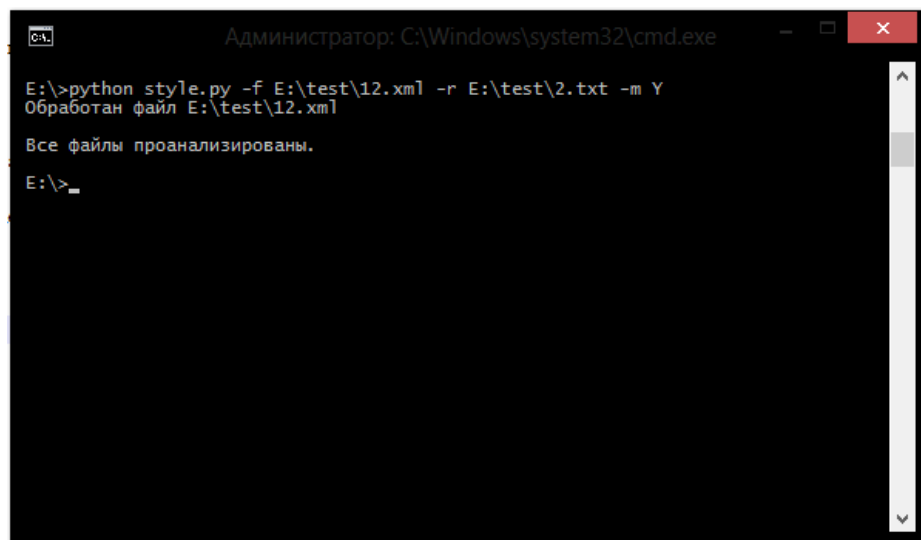
E:\>
```

Рис. 2. Пример запуска утилиты с параметром «Вызов справки»

После того, как пользователь ввёл необходимые параметры и запустил программу, в консоль последовательно выводятся полные пути к проанализированным файлам. После завершения обработки всех файлов в консоль выводится сообщение «Все файлы проанализированы», и программа

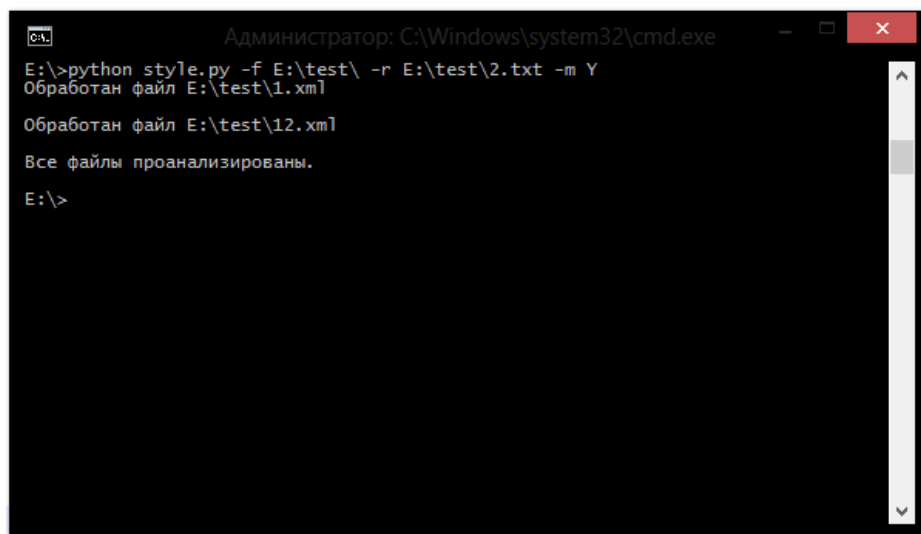


прекращает работу. Пример запуска утилиты с пользовательскими параметрами приведён на рисунках 3а (запуск программы для обработки отдельного файла) и 3б (запуск программы для обработки всех файлов, находящихся в определённой папке).



```
Администратор: C:\Windows\system32\cmd.exe
E:\>python style.py -f E:\test\12.xml -r E:\test\2.txt -m Y
Обработан файл E:\test\12.xml
Все файлы проанализированы.
E:\>_
```

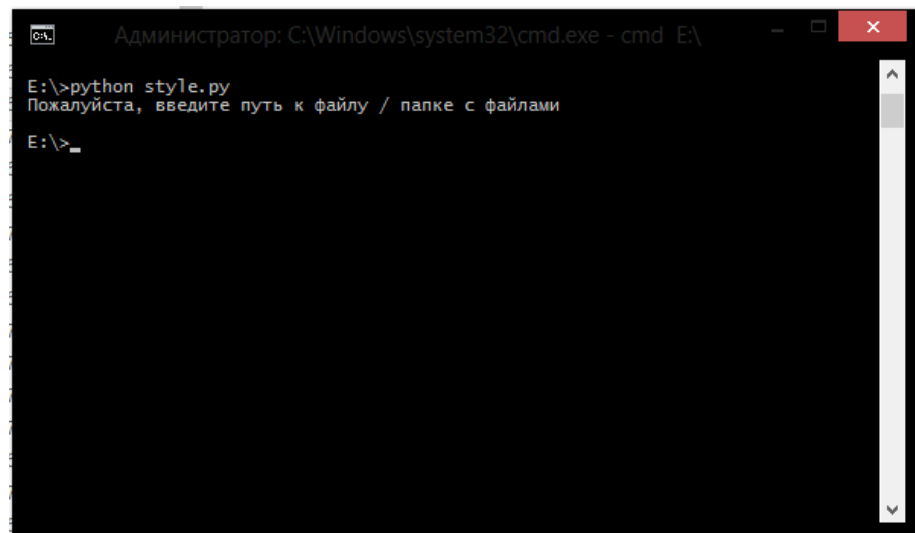
*Рис. 3а. Пример запуска утилиты с пользовательскими параметрами (обработка определённого файла)*



```
Администратор: C:\Windows\system32\cmd.exe
E:\>python style.py -f E:\test\ -r E:\test\2.txt -m Y
Обработан файл E:\test\1.xml
Обработан файл E:\test\12.xml
Все файлы проанализированы.
E:\>
```

*Рис. 3б. Пример запуска утилиты с пользовательскими параметрами (обработка всех файлов в определённой папке)*

Если пользователь запустит программу, не передав никакие параметры, или введёт несуществующий путь к файлу или папке программа выведет в консоль сообщение «Пожалуйста, введите путь к файлу / папке с файлами» и не начнёт работу (рис. 4).



*Рис. 4. Пример запуска утилиты без введения параметров*

### **ГЛАВА 3. ЭКСПЕРИМЕНТАЛЬНАЯ ПРОВЕРКА ВОЗМОЖНОСТИ АВТОМАТИЧЕСКОЙ СТИЛИСТИЧЕСКОЙ КЛАССИФИКАЦИИ РУССКОЯЗЫЧНЫХ ТЕКСТОВ**

#### **3.1. Подготовка корпусов**

Для создания экспериментальных корпусов нами были отобраны похожие по лексическому составу тексты четырёх функциональных стилей:

- 1) Научный стиль – тексты по радиоэлектронике, ракетостроению и технике.
- 2) Художественный стиль – тексты – научно–фантастические произведения второй половины XX – начала XXI века.
- 3) Деловой стиль – федеральные государственные образовательные стандарты высшего профессионального образования (ФГОС) по направлениям подготовки (специальностям), связанным с радиоэлектроникой, ракетостроением и техникой («Астрономия», «Радиотехника», «Космонавтика» и пр.), а также рабочие программы по дисциплинам, изучаемым в вузах по данным специальностям («Введение в авиационную и ракетно–космическую технику», «Основы ракетных двигателей» и пр.) за 2000–2015 гг.
- 4) Публицистический стиль – статьи из выпусков журналов «Новости космонавтики» и «Аэрокосмическая техника» за 1989–2000 гг.

Список всех использованных текстов приведён в Приложении Б.

#### **3.2. Подбор характеризующих признаков**

Выбор исходного набора параметров – одна из центральных задач проблемы классификации. В функциональной стилистике к настоящему времени накоплен уже значительный материал по параметризации стилей [Марусенко, 1990, с. 67].

Для определения стилевой принадлежности текстов нами были использованы индексы, не требующие информации о морфемной разметке. Опираясь на описания характеристик стилей из пунктов 1.1.1.–1.1.4. нашей работы, мы приняли за основу индексы, предложенные А.Ф. Журавлевым в работе «Опыт количественно–типологического исследования разновидностей устной речи» [Журавлёв, 1988, с. 84–150], и индексы, предложенные М.А. Марусенко в работе «Атрибуция анонимных и псевдонимных литературных произведений методами теории распознавания образов» [Марусенко, 1990, с. 72–74].

Мы реализовали подсчёт следующих параметров для текстов исследуемых корпусов:

1. Глагольность: отношение числа глаголов к числу слов в тексте;
2. Субстантивность: отношение числа существительных к числу словоформ в тексте;
3. Адъективность: отношение числа прилагательных к числу словоформ в тексте;
4. Отношение числа личных местоимений к числу словоформ в тексте;
5. Отношение числа частиц к числу словоформ в тексте;
6. Отношение числа междометий к числу словоформ в тексте;
7. Количество конструкций «существительное + существительное» (в том числе количество конструкций «существительное + существительное в родительном падеже»);
8. Количество конструкций «глагол + существительное»;
9. Комбинированный параметр, отражающий соотношение динамичности / статичности текстов коллекции (подробнее о данном параметре см. работу [Антонова, Клышинский, Ягунова, 2011]);
10. Средняя длина слова (число символов от пробела до пробела);
11. Средняя длина предложения (от точки до точки).

Следует отметить, что об устойчивом различии текстов по указанным параметрам говорят, в частности, А.Г. Поспелова и Е.В. Ягунова в своих исследованиях [Поспелова, Ягунова, 2014a] и [Поспелова, Ягунова, 2014b], а также ряд других учёных (см., например, работу [Клышинский и др., 2013]). Также на возможность стилистической дифференциации текстов по морфологическим параметрам указывает П.И. Браславский в статье [Браславский, 2003].

### **3.3. Ход экспериментов**

Собранные нами корпуса были обработаны гибридным морфологическим анализатором для русского языка на основе NLTK и PyMorphy2 – NLTK4RUSSIAN (см. работу [Паничева и др., 2015]). Результат работы анализатора представляет собой xml-файл, содержащий морфологическую и лексико-грамматическую разметку всех слов, входящих в выбранный корпус. Пример разметки конкретного слова можно увидеть в п. 2.2 нашей работы.

Для определения релевантных статистических параметров в составе корпусов нами были выделены подкорпусы, состоящие из 35 текстов каждый. Объём текстов составлял от 10 до 20 тыс. словоупотреблений. Обработав эти тексты при помощи статистического модуля нашей программы, мы получили возможность сравнить количественные характеристики текстов различных функциональных стилей и определить, какие из них являются характеризующими для каждого из них.

Целью анализа являлась проверка возможности классификации текстов по их принадлежности к разным функциональным стилям (научному, художественному, деловому и публицистическому) при помощи подобранных частотных характеристик.

### 3.4. Анализ данных

#### 3.4.1. Анализ лексико–морфологических индексов

Исследование лексико–морфологических параметров проводилось внутри подкорпусов отдельно для каждого текста; затем проводились вычисления средних значений каждого параметра, а также выявление минимальных и максимальных их значений для всей совокупности документов.

В таблицах 1–5 представлено распределение слов различных частей речи по подкорпусам: *min* – минимальное значение, *max* – максимальное значение, *mean* – среднее значение, *StD* – стандартное отклонение. Наибольшие значения выделены полужирным шрифтом, наименьшие – курсивом.

*Таблица 1. Распределение имён существительных по подкорпусам*

| Параметр    | Художественный стиль | Научный стиль | Публицистический стиль | Деловой стиль |
|-------------|----------------------|---------------|------------------------|---------------|
| <b>min</b>  | <i>0,2264</i>        | 0,3394        | 0,2949                 | <b>0,4165</b> |
| <b>max</b>  | <i>0,3267</i>        | 0,4552        | 0,3859                 | <b>0,5152</b> |
| <b>mean</b> | <i>0,2763</i>        | 0,392         | 0,3583                 | <b>0,4504</b> |
| <b>StD</b>  | <b>0,257</b>         | 0,0238        | <i>0,0182</i>          | 0,0193        |

Согласно данным, представленным в таблице 1, наибольшее число существительных содержится в текстах научного и делового стиля. Это объясняется тем, что для текстов этих стилей характерна бóльшая статичность: они не описывают происходящее событие, а констатируют факт его существования. В научных текстах часто вместо сказуемого, выраженного формой глагола, используется конструкция, состоящая из отглагольного существительного и глагола с ослабленным лексическим значением (напр., *наблюдается незначительное повышение температуры, ожидается повышение атмосферного давления*). Для текстов, относящихся к

деловому стилю, также характерно употребление большого количества отглагольных существительных (*несоблюдение, выполнение* и т.д.).

*Таблица 2. Распределение имён прилагательных по подкорпусам*

| Параметр    | Художественный<br>стиль | Научный стиль | Публицистический<br>стиль | Деловой<br>стиль |
|-------------|-------------------------|---------------|---------------------------|------------------|
| <b>min</b>  | 0,0709                  | 0,1152        | 0,0941                    | <b>0,14</b>      |
| <b>max</b>  | 0,1436                  | 0,198         | 0,1474                    | <b>0,217</b>     |
| <b>mean</b> | 0,1155                  | 0,1512        | 0,1217                    | <b>0,196</b>     |
| <b>StD</b>  | 0,0155                  | <b>0,0195</b> | 0,0135                    | 0,0139           |

Из таблицы 2 следует, что наибольшее число имён прилагательных содержится в текстах делового и научного стиля. Частотность прилагательных в текстах этих стилей обусловлена их включением в составные термины (напр., *магнитное поле, учебная программа*).

Небольшое число прилагательных в текстах художественного стиля, вероятно, обусловлено спецификой выбранной тематики: исследованные нами тексты принадлежат к жанру так называемой твёрдой научной фантастики. В текстах этого жанра особое внимание уделяется описанию открытий и научно–технических изобретений, и обилие качественных прилагательных, характерное для художественных текстов других жанров, им не свойственно. (более подробно о твёрдой научной фантастике см., например, работу [9]).

*Таблица 3. Распределение глаголов по подкорпусам*

| Параметр    | Художественный<br>стиль | Научный стиль | Публицистический<br>стиль | Деловой<br>стиль |
|-------------|-------------------------|---------------|---------------------------|------------------|
| <b>min</b>  | <b>0,1473</b>           | 0,0493        | 0,0779                    | 0,0208           |
| <b>max</b>  | <b>0,1983</b>           | 0,1013        | 0,1252                    | 0,0613           |
| <b>mean</b> | <b>0,167</b>            | 0,0791        | 0,0962                    | 0,0505           |
| <b>StD</b>  | <b>0,0138</b>           | 0,0135        | 0,0129                    | 0,0086           |

Согласно таблице, наибольшее число глаголов содержится в текстах, принадлежащих к художественному стилю. Стоит отметить, что тексты этого стиля отличает динамичность: в них реализуется большое количество ситуаций – следовательно, в них используется большее количество личных и неличных форм глагола. Тексты научного и делового стилей, как уже указывалось выше, наоборот, отличает статичность и номинативность.

*Таблица 4. Распределение местоимений по подкорпусам*

| Параметр    | Художественный стиль | Научный стиль | Публицистический стиль | Деловой стиль |
|-------------|----------------------|---------------|------------------------|---------------|
| <b>min</b>  | <b>0,0446</b>        | 0,0038        | 0,0087                 | 0,0024        |
| <b>max</b>  | <b>0,1117</b>        | 0,0231        | 0,0446                 | 0,0113        |
| <b>mean</b> | <b>0,0695</b>        | 0,0124        | 0,0183                 | 0,0065        |
| <b>StD</b>  | <b>0,0159</b>        | 0,004         | 0,0077                 | 0,0018        |

Согласно полученным данным, наибольшее количество местоимений содержится в текстах художественного стиля. Предположительно это связано с тем, что к текстам художественного стиля предъявляются особые требования в отношении отсутствия лексических повторов: подобные повторы в них возможны лишь в качестве специального средства речевой выразительности. В подавляющем большинстве случаев в художественных текстах используются синонимические ряды и замена существительных местоимениями. Также использование местоимений изредка может привести к двусмысленности: например, в предложении *Сестра поступила в артистическую труппу, и она отправляется на гастроли* неясно, какое существительное заменяет местоимение *она*. Подобная неопределённость недопустима в текстах научного и делового стиля поэтому употребление в них местоимений сведено к минимуму.



Таблица 5. Распределение частиц по подкорпусам

| Параметр    | Художественный стиль | Научный стиль | Публицистический стиль | Деловой стиль |
|-------------|----------------------|---------------|------------------------|---------------|
| <b>min</b>  | <b>0,0232</b>        | 0,0032        | 0,0089                 | <i>0,0001</i> |
| <b>max</b>  | <b>0,0673</b>        | 0,0284        | 0,0278                 | <i>0,0064</i> |
| <b>mean</b> | <b>0,0491</b>        | 0,0118        | 0,0169                 | <i>0,0026</i> |
| <b>StD</b>  | <b>0,0105</b>        | 0,0051        | 0,005                  | <i>0,0017</i> |

Полученные данные свидетельствуют о том, что наибольшее число частиц содержится в текстах художественного стиля, а наименьшее – в текстах делового стиля. Это подтверждает выделенные нами ранее характерные черты этих стилей речи: для художественных текстов характерно использование разнообразных лексических средств, в том числе и частиц, многие из которых вносят в предложения эмоциональные оттенки (напр., сомнение: *вряд ли, едва ли*; удивление: *что за, как* и т.д.); для делового стиля, напротив, характерно почти полное отсутствие эмоционально–экспрессивных речевых средств.

Обратим внимание на то, что значения лексических параметров художественного и делового стиля находятся на разных концах шкалы: если значения параметра художественного текста максимальны, то значения этого же параметра делового текста минимальны, и наоборот.

Также следует отметить, что значения всех описанных параметров публицистического стиля всегда находятся в промежутке между значениями соответствующих параметров научного и художественного стилей. Это связано с тем, что, как отмечалось ранее, для публицистического функционального стиля характерно совмещение характеристик научного (использование терминов) и художественного стилей (динамичность, использование эмоционально–экспрессивных речевых средств), что делает задачу правильной автоматической классификации текстов, принадлежащих к нему, более трудоёмкой.

### **3.4.2. Анализ материала на основе данных о частеречной сочетаемости**

Метод анализа синтаксических конструкций текста на основе частеречной сочетаемости был использован для анализа особенностей синтаксической структуры текстов разных функциональных стилей, жанров и/или предметных областей в работах [Клышинский и др., 2013] и [Антонова, Клышинский, Ягунова, 2011]. Этот подход, использующий статистический метод извлечения информации о частеречной сочетаемости слов, показал свою эффективность для определения функционального стиля коллекций текстов.

Само по себе понятие конструкции довольно широкое и не вполне конкретное: согласно общепринятому мнению, конструкция есть некое языковое выражение, сформированное из фиксированного компонента (например, целевого слова) и слотов, заполняемых контекстными соседями с теми или иными лексическими, грамматическими и другими признаками. Поскольку до конца не выяснено, какие именно типы словосочетаний могут служить показателями для чёткого разделения текстов по стилям, для решения нашей исследовательской задачи мы приняли решение воспользоваться понятием контекстного профиля целевой леммы (подробнее о контекстных профилях см. [Верёвкина и др., 2013; Ляшевская и др., 2012; Митрофанова и др., 2010]).

В таблицах 6–7 представлено распределение конструкций «существительное + существительное» и «существительное + существительное в родительном падеже» по подкорпусам: min – минимальное значение, max – максимальное значение, mean – среднее значение, StD – стандартное отклонение. Наибольшие значения выделены полужирным, наименьшие – курсивом.

*Таблица 6. Распределение конструкций «существительное + существительное» по подкорпусам*

| Параметр    | Художественный стиль | Научный стиль | Публицистический стиль | Деловой стиль  |
|-------------|----------------------|---------------|------------------------|----------------|
| <b>min</b>  | 193                  | 690           | 552                    | <b>941</b>     |
| <b>max</b>  | 498                  | 1304          | 1035                   | <b>1496</b>    |
| <b>mean</b> | 349,7                | 1038,15       | 845,59                 | <b>1179,26</b> |
| <b>StD</b>  | 89,05                | <b>150,22</b> | 91,76                  | 144,143        |

*Таблица 7. Распределение конструкций «существительное + существительное в родительном падеже» по подкорпусам*

| Параметр    | Художественный стиль | Научный стиль | Публицистический стиль | Деловой стиль |
|-------------|----------------------|---------------|------------------------|---------------|
| <b>min</b>  | 107                  | 535           | 350                    | <b>816</b>    |
| <b>max</b>  | 357                  | 1105          | 705                    | <b>1393</b>   |
| <b>mean</b> | 229,62               | 863,41        | 587,15                 | <b>1015</b>   |
| <b>StD</b>  | 69,67                | <b>145,87</b> | 72,31                  | 131,44        |

Таблицы показывают, что конструкции данных типов преобладают в текстах научного и делового стиля. Это объясняется, в частности, высокой номинативностью текстов, принадлежащих к этим функциональным стилям.

Ранее мы также указывали, что одна из отличительных особенностей текстов делового стиля – использование цепочек родительных падежей (см. п. 1.1.4.). Высокие значения этого параметра для текстов научного стиля объясняются тем, что конструкция «существительное + существительное в родительном падеже» – это генитивная конструкция, часто соответствующая неоднословному термину (напр., *скорость света*, *система координат*). Большое количество этих конструкций традиционно рассматривается как морфо–синтаксическая характеристика научных текстов.

В таблице 8 представлено распределение конструкций «глагол + существительное» по подкорпусам: min – минимальное значение, max –

максимальное значение, mean – среднее значение, StD – стандартное отклонение. Наибольшие значения выделены полужирным, наименьшие – курсивом.

*Таблица 8. Распределение конструкций «глагол + существительное» по подкорпусам*

| Параметр    | Художественный стиль | Научный стиль | Публицистический стиль | Деловой стиль |
|-------------|----------------------|---------------|------------------------|---------------|
| <b>min</b>  | <b>313</b>           | 232           | 304                    | <i>11</i>     |
| <b>max</b>  | <b>646</b>           | 539           | 560                    | <i>173</i>    |
| <b>mean</b> | <b>466,85</b>        | 369,32        | 430,03                 | <i>96,56</i>  |
| <b>StD</b>  | 77,31                | <b>88,94</b>  | 66,28                  | <i>36,56</i>  |

Большое количество конструкций «глагол + существительное» характеризует тексты, в которых реализуется большое число ситуаций (т.е. динамические тексты). Обилие же конструкций «существительное + существительное», наоборот, отличает статические тексты (т.е. тексты, в которых описывается некоторое положение дел). Таким образом, можно сказать, что тексты с большим количеством конструкций «глагол + существительное» описывают какие-то события, происшествия, а тексты с маленьким числом этих конструкций – указывают на наличие неких событий, называют их.

В п. 3.4.1. при исследовании лексических параметров текстов различных стилей мы уже отмечали, что художественные тексты, в отличие от научных и деловых, характеризуются бóльшим числом глаголов (т.е. большей динамичностью). На уровне конструкций это различие сохраняется: данные, представленные в таблице 8, указывают на то, что наибольшее число конструкций «глагол + существительное» содержится именно в художественных текстах.

Следует опять обратить внимание на то, что и на уровне конструкций значения параметров текстов публицистического стиля находятся в

промежутке между значениями индексов художественного и научного стилей.

В работе [3] был определён комбинированный параметр  $\beta$ , отражающий соотношение динамичности и статичности текстов коллекции. Этот параметр выглядит следующим образом:

$$\beta = \frac{\#(\text{гл} + \text{сущ}) + \#(\text{гл} + \text{нар}) + \#(\text{деепр} + \text{сущ}) + \#(\text{деепр} + \text{нар})}{\#(\text{сущ} + \text{сущ}) + \#(\text{прил} + \text{сущ})},$$

где # – количество конструкций определённого типа. Очевидно, что в числителе используются показатели количества конструкций, свидетельствующих о динамичности текста (конструкции с глаголами и деепричастиями), а в знаменателе – показатели количества конструкций, свидетельствующих о статичности текста (конструкции с существительными).

Мы реализовали подсчёт данного параметра для текстов, входящих в наши подкорпуса. Результаты представлены в таблице 9: min – минимальное значение, max – максимальное значение, mean – среднее значение, StD – стандартное отклонение. Наибольшие значения выделены полужирным, наименьшие – курсивом.

*Таблица 9. Распределение значений параметра  $\beta$  по подкорпусам*

| Параметр    | Художественный стиль | Научный стиль | Публицистический стиль | Деловой стиль |
|-------------|----------------------|---------------|------------------------|---------------|
| <b>min</b>  | <b>0,4813</b>        | 0,13          | 0,2451                 | <i>0,0038</i> |
| <b>max</b>  | <b>1,2575</b>        | 0,3195        | 0,4963                 | <i>0,0687</i> |
| <b>mean</b> | <b>0,7844</b>        | 0,216         | 0,3279                 | <i>0,0388</i> |
| <b>StD</b>  | <b>0,1712</b>        | 0,0546        | 0,0679                 | <i>0,0157</i> |

Очевидно противопоставление деловых и художественных текстов по данному параметру: деловые показывают наименьшие значения, художественные – наивысшие. Следует также отметить, что промежутки, в

которые попадает данный параметр, не пересекаются для текстов почти всех стилей. Исключение составляют публицистические тексты: для них минимальные значения данного параметра находятся в диапазоне, соответствующем значениям данного показателя для научных текстов ( $0,25 \in [0,13; 0,32]$ ), а максимальные – в диапазоне, соответствующем значениям данного показателя для художественных текстов ( $0,49 \in [0,48; 1,26]$ ).

Полученные данные свидетельствуют о возможности использования данного параметра для определения стиля исследуемого текста. К такому же выводу приходят и авторы работы [3].

### 3.4.3. Параметры длины слова и длины предложения

В таблицах 10–11 представлено распределение длин слов (от пробела до пробела) и длин предложений (от точки до точки) по подкорпусам: *min* – минимальное значение, *max* – максимальное значение, *mean* – среднее значение, *StD* – стандартное отклонение. Наибольшие значения выделены полужирным, наименьшие – курсивом.

*Таблица 10. Распределение длин слов по подкорпусам*

| Параметр    | Художественный стиль | Научный стиль | Публицистический стиль | Деловой стиль |
|-------------|----------------------|---------------|------------------------|---------------|
| <b>min</b>  | 1                    | 1             | 1                      | 1             |
| <b>max</b>  | <b>40</b>            | <i>34</i>     | <i>37</i>              | <i>39</i>     |
| <b>mean</b> | <i>5,64</i>          | <i>6,75</i>   | <i>6,33</i>            | <b>7,99</b>   |
| <b>StD</b>  | <i>3,24</i>          | <i>3,85</i>   | <i>3,73</i>            | <b>4,5</b>    |

Заметим, что минимальное значение параметра длины слова для всех текстов равно 1. Во-первых, это связано с использованием в текстах всех стилей однобуквенных предлогов и союзов (*в, а, и* и пр.). Также однобуквенными являются, например, сокращения *т* (*тонна*), *г* (*год / грамм*) и др., то есть слова, имеющие семантику «мера» (масса, продолжительность

и др.) – очевидно, что подобные слова также встречаются в текстах разных стилей, но особенно их много в текстах научного стиля.

Рассмотрим слова из каждого подкорпуса, имеющие максимальную длину:

- Художественный стиль – *Хронально–гравитационно–пространственный* (40 символов);
- Научный стиль – *экспериментально–производственный* (34 символа);
- Публицистический стиль – *глицеральдегид–3–фосфат–дегидрогеназа* (37 символов);
- Деловой стиль – *проектно–конструкторско–технологическая* (39 символов).

Отметим, что слово, использованное в художественном тексте, создано автором текста (в результатах поиска, произведённого системой google по запросу «Хронально–гравитационно–пространственный», нет ни одного полного совпадения с данным сочетанием), а слово из публицистического текста является названием химического соединения, то есть терминологическим элементом (вспомним, что публицистический стиль может включать в себя, например, термины из различных научных областей). Очевидно, что максимальное и минимальное значение данного индекса не могут являться характеризующими параметрами, позволяющими однозначно отнести текст к определённому функциональному стилю. Мы предполагаем, однако, что при классификации документов можно воспользоваться параметром «средняя длина слова в тексте» как вспомогательным.

В таблице 11 представлено распределение длин предложений по подстилям. Как и в случае с параметром длины слова, наименьшее значение данного параметра для всех стилей равно 1. Очевидно, что это связано с использованием во всех стилях односоставных нераспространённых предложений, например: *Тишина.* (для художественных текстов); *Внимание!*

(для публицистических текстов); *Приложение*. (для научных и деловых текстов).

*Таблица 11. Распределение длин предложений по подкорпусам*

| Параметр    | Художественный стиль | Научный стиль | Публицистический стиль | Деловой стиль |
|-------------|----------------------|---------------|------------------------|---------------|
| <b>min</b>  | 1                    | 1             | 1                      | 1             |
| <b>max</b>  | 95                   | 174           | 262                    | <b>3130</b>   |
| <b>mean</b> | 9,39                 | 16,62         | 16,27                  | <b>44,6</b>   |
| <b>StD</b>  | 7,28                 | 10,83         | 11,5                   | <b>139,48</b> |

Наибольшее значение данного параметра наблюдается в предложениях делового стиля. Отличительными особенностями текстов данного стиля, как указывалось в п. 1.1.4., является тенденция к употреблению сложноподчинённых предложений, отражающих логическое подчинение одних фактов другим, а также использование номинативных предложений с перечислением. Именно эти особенности обуславливают столь высокие значения индекса длины предложения.

### **3.5. Инструмент автоматического определения стилистической принадлежности текстов**

На основании выявленных нами характеризующих параметров текстов различных функциональных стилей речи нами был разработан программный модуль, позволяющий провести автоматическую стилевую диагностику исследуемого текста. Рассмотрим подробно алгоритм его работы.

#### **3.5.1. Описание алгоритма стилистической принадлежности текстов**

Для построения инструмента автоматического определения стилистической принадлежности текстов нами было принято решение использовать метод деревьев принятия решений. Предполагается, что, последовательно сравнивая статистические параметры текста с



пороговыми значениями, полученными после анализа результатов проведённых экспериментов, можно пройти по дереву и определить стиль обрабатываемого текста.

Результаты анализа экспериментов позволяют утверждать, что тексты, относящиеся к деловому стилю, можно отделить от текстов других стилей с наибольшей точностью. Об этом свидетельствуют как результаты исследования лексико–морфологических индексов (тексты делового стиля имеют наивысшие значения индексов отношения числа существительных, и прилагательных к общему числу слов в тексте, и наименьшие значения индексов отношения числа глаголов, частиц и личных местоимений к общему числу слов в тексте), так и результаты исследования материала на основе частеречной сочетаемости (наибольшее число групп «существительное + существительное» и наименьшее – «глагол + существительное»).

Хуже всего отделяются тексты публицистического стиля: как многократно отмечалось ранее, значения параметров текстов этого стиля занимают промежуточное положение между значениями параметров художественного и научного стиля.

Сначала на основании параметров «средняя длина предложения» и «соотношение динамичности и статичности» проверяется, относится ли обрабатываемый текст к деловому стилю. Если значения параметров данного текста выходят за указанные рамки, осуществляется проверка принадлежности текста к художественному стилю (на основании параметров «доля частиц в тексте» и «доля местоимений в тексте»). Если значения данных параметров также выходят за пределы указанных диапазонов, осуществляется проверка параметров «доля прилагательных в тексте», «доля частиц в тексте» (данный параметр проверяется повторно, на данном этапе проверяется верхняя граница его значения) и «соотношение динамичности и статичности» (данный параметр также проверяется повторно, его верхняя граница повышается).

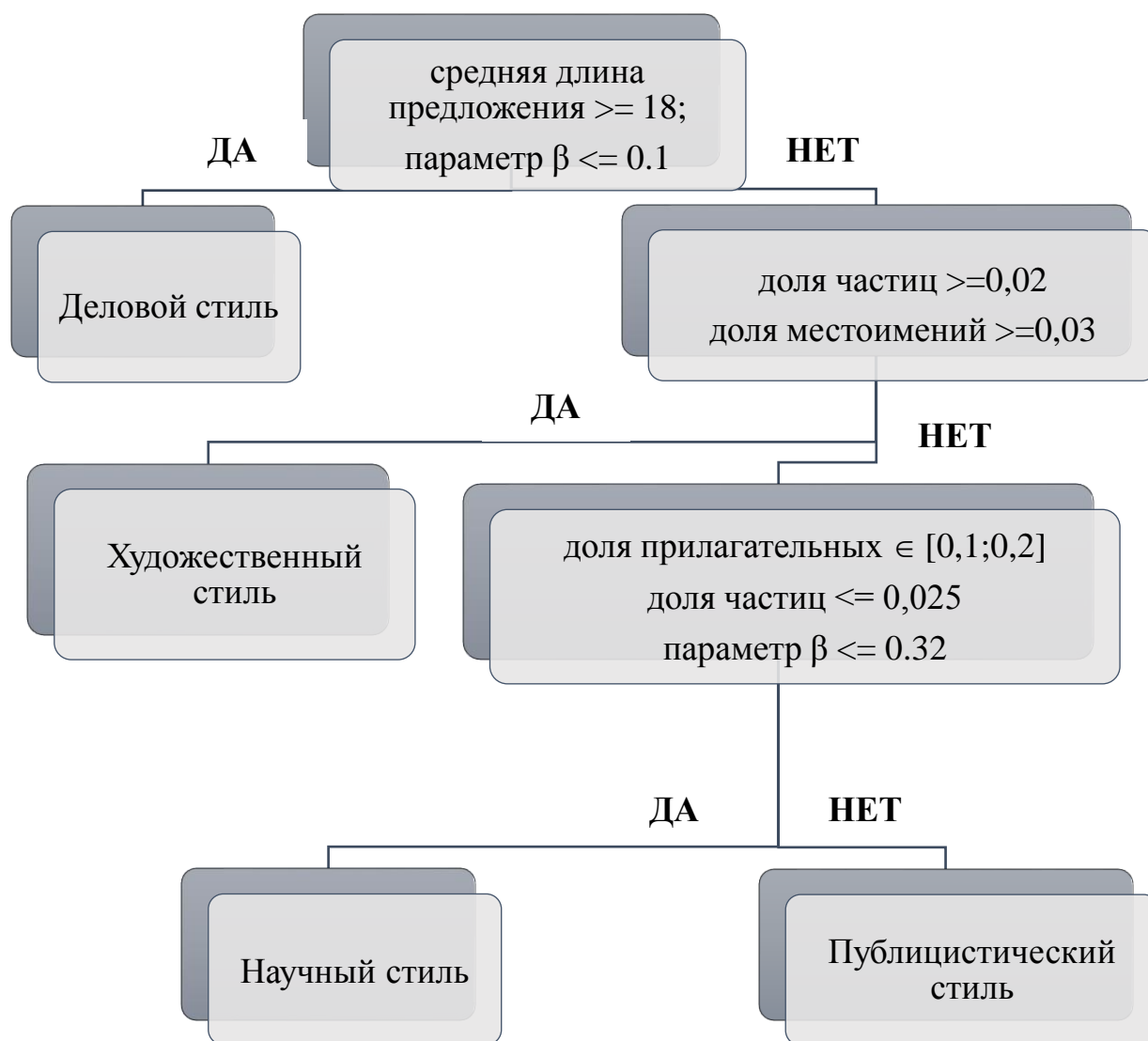


Рис. 2. Алгоритм работы модуля определения стилистической принадлежности текста

Разработанный нами алгоритм определения стилистической принадлежности исследуемого текста можно представить в виде дерева, изображённого на рис. 2.

Модуль определения стилистической принадлежности текста начинает работу после того, как статистическая обработка исходных параметров текста завершена. Программа поочерёдно сравнивает статистические параметры с пороговыми значениями, полученными в результате

произведённых экспериментов, и на основании этой оценки выдаёт предположение о стилистической принадлежности исследуемого текста.

С полным кодом программы можно ознакомиться в приложении А.

### **3.5.2. Оценка качества работы модуля автоматического определения стилистической принадлежности текстов**

С помощью разработанной нами программы мы осуществили стилистическую диагностику оставшихся текстов из собранных нами корпусов. Для каждого функционального стиля было проанализировано по 65 текстов (всего было обработано 260 текстов). Результаты работы утилиты представлены в таблице 12.

*Таблица 12. Результаты работы инструмента автоматического определения стилистической принадлежности текстов*

| <b>Стиль текста</b>     | <b>Кол–во текстов, для которых стиль определён верно</b> | <b>Кол–во текстов, для которых стиль определён не верно</b> |
|-------------------------|--|---|
| <b>Деловой</b>          | 65   | 0   |
| <b>Художественный</b>   | 65   | 0   |
| <b>Научный</b>          | 62   | 3   |
| <b>Публицистический</b> | 37   | 28  |

Проанализируем полученные результаты.

Деловые и художественные тексты программа безошибочно отнесла к соответствующим стилям. Очевидно, что параметры, подобранные нами (средняя длина предложения и соотношение динамичности и статичности), действительно являются характеризующими для текстов данных стилей и позволяют производить их классификацию с большой точностью.

С меньшей точностью была произведена классификация текстов научного стиля: три из шестидесяти пяти текстов научных текстов были отнесены утилитой к публицистическому стилю. Стоит, однако, отметить, что данные тексты относятся к научно–популярному подстилю – одной из

его особенностей является, например, использование экспрессивных средств выразительности при сохранении характерной для научных текстов чёткости изложения (подробнее о научно–популярном подстиле см. в работе [Хомутова, Петров, 2013]). Можно сказать, что научно–популярные тексты находятся на стыке научного и публицистического стилей.

Больше всего ошибок было выявлено при классификации публицистических текстов. Из 65 обработанных текстов лишь 37 (57%) было правильно отнесено к публицистическому стилю. Оставшиеся тексты были классифицированы следующим образом:

- 23 текста были отнесены к научному стилю;
- 5 текстов были отнесены к художественному стилю.

Следует обратить внимание на то, что тексты, отнесённые программой к художественному стилю, написаны в форме бортового журнала космонавтов и относятся к жанру «художественной публицистики». Тексты этого жанра сближаются по своим характеристикам с научными текстами, сочетают функцию привлечения внимания адресата к описываемому явлению и эстетическую функцию (подробнее о жанре «художественной публицистики» см. в работе [Прохоров, 2012]).

Тексты, отнесённые к научному стилю, характеризуются большим количеством терминов. В данных текстах количество сочетаний «существительное + существительное» (генитивная конструкция, часто соответствующая неоднословному термину) значительно превышает среднее значение для текстов публицистического стиля, полученное нами в результате экспериментов. Увеличение этого параметра нарушает работу алгоритма (значение параметра  $\beta$  понижается, что свидетельствует о превалировании статичности над динамичностью), и программа классифицирует такие тексты как научные.

Произведём оценку работы нашего классификатора для каждого стиля (таблица 13).

*Таблица 13. Оценка работы классификатора*

| <b>Стиль текста</b>     | <b>Точность</b> | <b>Полнота</b> | <b>F–мера</b> |
|-------------------------|-----------------|----------------|---------------|
| <b>Деловой</b>          | 1               | 1              | 1             |
| <b>Художественный</b>   | 0,99            | 1              | 0,99          |
| <b>Научный</b>          | 0,73            | 0,95           | 0,83          |
| <b>Публицистический</b> | 0,93            | 0,57           | 0,7           |

В целом можно утверждать, что разработанный нами инструмент успешно справился с задачей классификации текстов различных стилей. Представляется возможным улучшить результаты классификации текстов, принадлежащих к публицистическому и научному стилям, используя параметры более высокого уровня (например, синтаксические). Также возможно сделать классификацию более подробной, добавив разделение на подстили и жанры и подобрав характеризующие параметры для каждого из них.

## **ЗАКЛЮЧЕНИЕ**

В данной работе мы подробно изучили вопрос о выделении различных функциональных стилей в современном русском языке и описали основные методы автоматической классификации текстов. Нами были выявлены характерные особенности четырёх стилей русского языка – научного, официально–делового, художественного и публицистического – и выдвинута гипотеза о том, что возможно подобрать такие комбинации параметров, которые позволят однозначно определять стиль исследуемого текста.

Сравнив коллекции текстов, принадлежащих к вышеуказанным функциональным стилям, при помощи разработанного нами модуля статистической обработки текстов, мы выделили параметры, позволяющие наиболее точно разграничить документы, относящиеся к разным стилям. Эти индексы легли в основу разработанного нами инструмента автоматического определения стилистической принадлежности текстов. Проанализировав при помощи данного инструмента по 65 текстов из собранных нами корпусов, мы успешно классифицировали более 88% из них, причём наибольшая точность была достигнута при классификации деловых и художественных текстов. Это подтвердило наше первоначальное предположение о возможности автоматической классификации документов, относящихся к разным функциональным стилям.

В дальнейшем представляется возможным изучить большее число статистических характеристик отдельных текстов или их фрагментов, а также усложнить параметры, используемые при классификации текстов.

**Перспективы развития** нашего исследования связаны:

- 1) с усложнением и совершенствованием разработанного нами инструмента: например, за счёт использования большего числа параметров разных типов (синтаксических, морфологических и др.) отдельно, а также в комбинации с уже изученными индексами.

- 2) с расширением экспериментального материала и проведением исследований по автоматической обработке бóльшего числа корпусов текстов из других коллекций (например, текстов разговорного стиля или текстов, относящихся к различным литературным жанрам).

## СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Азимов Э. Г., Щукин А. Н. Новый словарь методических терминов и понятий (теория и практика обучения языкам). – М.: Издательство ИКАР, 2009. 448 с.

2. Андреев В.С. Классификация стихотворных текстов (на материале лирики американских поэтов романтиков): автореф. дис. ... канд. филол. наук. Смоленск, 2002.

3. Антонова А.Ю., Клышинский Э.С., Ягунова Е.В. Определение стилевых и жанровых характеристик коллекций текстов на основе частеречной сочетаемости // Труды международной конференции «Корпусная лингвистика–2011». – СПб.: С.–Петербургский гос. университет, Филологический факультет, 2011

URL:

[http://webground.su/data/lit/antonova\\_klyshinsky\\_yagunova/Opredeleniye\\_stilevyh\\_i\\_zhanrovyyh\\_kharakteristik.pdf](http://webground.su/data/lit/antonova_klyshinsky_yagunova/Opredeleniye_stilevyh_i_zhanrovyyh_kharakteristik.pdf) (дата последнего обращения: 17.05.2017)

4. Бикмуканова С. И. Публицистический стиль и его функционирование // Science Time. 2014. №12 (12). С. 36–37

URL: <http://cyberleninka.ru/article/n/publitsisticheskiy-stil-i-ego-funktsionirovanie> (дата последнего обращения: 17.05.2017).

5. Березин Ф.М., Головин Б.Н. Общее языкознание. М.: Просвещение, 1979. 415 с.

6. Большаков А.А., Каримов Р.Н. Методы обработки многомерных данных и временных рядов. М.: Горячая линия – Телеком, 2007. 522 с.

7. Бондарь А. С., Каширина Н. А. Особенности перевода клише в текстах газетно–публицистического стиля // Символ науки. 2016. №2–3. С. 52–54

URL: <http://cyberleninka.ru/article/n/osobennosti-perevoda-klishhe-v-tekstah-gazetno-publitsisticheskogo-stilya> (дата последнего обращения: 17.05.2017).



8. Браславский П. Опыт автоматической классификации текстов по стилям (на материале документов Internet) // Русский язык в Интернете. Сб. статей. Казань, 2003. С. 6–15.

9. Бритиков А. Ф. Отечественная научно–фантастическая литература (1917–1991 годы). Книга вторая. Некоторые проблемы истории и теории жанра. Изд. 2–е, испр. и доп. СПб: Творческий центр «Борей–арт», 2005. 229 с.

10. Будагов Р.А. Литературные языки и языковые стили. М.: Высшая школа, 1967. 376 с.

11. Вартан А. Ю. Классификация ресурсов из сети Интернет по направлениям наркоторговля, терроризм, экстремизм // Вестник Югорского государственного университета. 2015. №S2 (37).

URL: <http://cyberleninka.ru/article/n/klassifikatsiya-resursov-iz-seti-internet-po-napravleniyam-narkotorgovlya-terrorizm-ekstremizm> (дата последнего обращения: 17.05.2017).

12. Васнецов А.Г. Сравнение эффективности некоторых статистических методов классификации на примере технических статей // Молодежный научно–технический вестник. 2015. №2.

13. Вережкина О.И., Донцова М.Д., Пушкина Т.А., Реброва П.В. Разработка и тестирование инструментов грамматического и лексико–семантического профилирования (на материале выборок из НКРЯ) // Материалы XXII международной филологической конференции. секция прикладной и математической лингвистики. СПб., 2013.

14. Виноградов В.В. Итоги обсуждения вопросов стилистики // Вопросы языкознания. – 1955. – № 1. С. 85;

15. Виноградов В.В. К теории литературных стилей (Виноградов В.В. Избранные труды. О языке художественной прозы. — М., 1980. С. 240–249)

URL: <http://philology.ru/linguistics2/vinogradov-80.htm> (дата последнего обращения: 17.05.2017)

16. Виноградов В. В. Стилистика. Теория поэтической речи. Поэтика. М., 1963.

17. Винокур. Культура языка. М., 1929

18. Винокур Г.О. Об изучении языка литературных произведений (Винокур Г.О. Избранные работы по русскому языку. — М., 1959. С. 229–259)

URL: <http://philology.ru/linguistics2/vinokur-59i.htm> (дата последнего обращения: 17.05.2017)

19. Воройский Ф.С. Информатика. Новый систематизированный толковый словарь–справочник. — 3–е изд.. — М.: ФИЗМАТЛИТ, 2003. — 760 с. — (Введение в современные информационные и телекоммуникационные технологии в терминах и фактах).

20. Воронцов К.В. Математические методы обучения по прецедентам (теория обучения машин) – 2012. — 160 с.

URL: <http://docplayer.ru/2064-K-v-voroncov-http-www-ccas-ru-voron-voron-ccas-ru.html> (дата последнего обращения: 17.05.2017)

21. Востоков А.Х. Русская грамматика. СПб, 1831. — 449 с.

URL: [https://books.google.ru/books?id=JDhAAAAAYAAJ&pg=PR7&hl=ru&source=gs\\_selected\\_pages&cad=2#v=onepage&q&f=false](https://books.google.ru/books?id=JDhAAAAAYAAJ&pg=PR7&hl=ru&source=gs_selected_pages&cad=2#v=onepage&q&f=false) (дата последнего обращения: 17.05.2017)

22. Головин Б.Н. Основы культуры речи. М., 1988. 320 с.

23. Горшков А.И. Русская стилистика. Стилистика текста и функциональная стилистика: учеб. для педагогических университетов и гуманитарных вузов / А.И. Горшков – М., АСТ: Астрель, 2006. 367 с.

24. Гулин В. В. Методы снижения размерности признакового описания документов в задаче классификации текстов. Вестник МЭИ №2 2013. С. 115–121.

25. Епрев А. С. Автоматическая классификация текстовых документов // МСМ. 2010. №1 (21).

URL: <http://cyberleninka.ru/article/n/avtomaticheskaya-klassifikatsiya-tekstovyh-dokumentov-1> (дата последнего обращения: 17.05.2017).

26. Ермолаева Ю. Е. Классификация стихотворных текстов методом дискриминантного анализа // Вестник ТГУ. 2009. №7.

URL: <http://cyberleninka.ru/article/n/klassifikatsiya-stihotvornyh-tekstov-metodom-diskriminantnogo-analiza> (дата последнего обращения: 17.05.2017).

27. Журавлев А. Ф. Опыт квантитативно–типологического исследования разновидностей устной речи // Разновидности городской устной речи: Сборник научных трудов. – М.: Наука, 1988.

28. Зайцева Т. В., Васина Н. В., Пусная О. П., Смородина Н. Н. Программная реализация метода деревьев решений для решения задач классификации и прогнозирования // Научные ведомости Белгородского государственного университета. Серия: Экономика. Информатика. 2013. №8–1 (151).

URL: <http://cyberleninka.ru/article/n/programmnyaya-realizatsiya-metoda-dereviev-resheniy-dlya-resheniya-zadach-klassifikatsii-i-prognozirovaniya> (дата последнего обращения: 17.05.2017).

29. Калмыков А. А., Коханова Л. А. Интернет–журналистика. М.: ЮНИТИ–ДАНА, 2005. 383 с.

30. Клышинский Э.С., Кочеткова Н.А., Мансурова О.Ю., Ягунова Е.В., Максимов В.Ю., Карпик О.В. Формирование модели сочетаемости слов русского языка и исследование ее свойств Москва // Препринты ИПМ им. М.В. Келдыша. 2013. № 41. 23 с.

31. Кожина М.Н. О соотношении стилей языка и стилей речи с позиций языка как функционирующей системы // Принципы функционирования языка в его речевых разновидностях. Пермь, 1984.

32. Кожина М. Н. Стилистика русского языка: учеб. для студентов пед. ин-тов по специальности "Рус. яз. и лит." / М. Н. Кожина. - Изд. 3-е., перераб. и доп. - М. : Просвещение, 1993. 223 с.

33. Левитин А.В. "Алгоритмы: введение в разработку и анализ" Вильямс, 2006. — С. 409–417. 576 с.
34. Лингвистический энциклопедический словарь / Под ред. В. Н. Ярцевой. – М.: Советская энциклопедия, 1990 [Электронный ресурс].  
URL: <http://tapemark.narod.ru/les/index.html> (дата последнего обращения: 17.05.2017)
35. Ломоносов М. В. Предисловие о пользе книг церковных в российском языке // Ломоносов М. В. Полн. собр. соч. — Т. 7. — М.; Л.: Изд-во АН СССР, 1952. С. 589—590.
36. Ляшевская О.Н., Митрофанова О.А., Грачкова М.А., Шиморина А.С., Шурыгина А.С., Романов С.В. К построению инвентаря русских именных конструкций // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая – 3 июня 2012г.). Вып. 11 (18). – М.: Изд-во РГГУ, 2012.
37. Марусенко М. А. Атрибуция анонимных и псевдонимных литературных произведений методами распознавания образов. – Л. : Изд-во Ленингр. ун-та, 1990. 164 с.
38. Митрофанова О.А., Грачкова М.А., Шиморина А.С., Ляшевская О.Н. Лексические, семантические и морфологические признаки контекстов в разрешении неоднозначности русских существительных // XXXIX Международная филологическая конференция. Секция математической лингвистики. СПб., 2010.
39. Москвин В.П. Стилистика русского языка. Теоретический курс. – Ростов–на–Дону: Феникс, 2006. 630 с.
40. Мурашова Л. П. Отечественная функциональная лингвистика // Научный вестник ЮИМ. 2017. №1.  
URL: <http://cyberleninka.ru/article/n/otechestvennaya-funktsionalnaya-lingvistika> (дата последнего обращения: 17.05.2017).
41. Одинцов В. В. Стилистика текста. М., 1980. С. 78.

42. Паничева П.В., Протопопова Е.В., Митрофанова О.А., Мирзагитова А.Р. Разработка лингвистического комплекса для морфологического анализа русскоязычных корпусов текстов на основе PyMorphy и NLTK // Труды международной конференции «Корпусная лингвистика – 2015». СПб., 2015. С. 361-373.

URL:

[http://mathling.phil.spbu.ru/sites/default/files/CORPORA2015\\_PyMorphy+NLTK\\_11.05.pdf](http://mathling.phil.spbu.ru/sites/default/files/CORPORA2015_PyMorphy+NLTK_11.05.pdf) (дата последнего обращения: 17.05.2017).

43. Панова Н.Ф., Денисова Н.В. Классификация студентов по уровню успеваемости с помощью аппарата дискриминантного анализа // Вестник ОГУ. 2014. №8 (169).

URL: <http://cyberleninka.ru/article/n/klassifikatsiya-studentov-po-urovnyu-uspevaemosti-s-pomoschyu-apparata-diskriminantnogo-analiza> (дата последнего обращения: 17.05.2017).

44. Поспелова А.Г., Ягунова Е.В.. Категоризация коллекций текстов на основе низкоуровневых параметров текста // Конференция AINL 2014: Искусственный интеллект и естественный язык

45. Поспелова А. Г., Ягунова Е. В. Опыт применения стилевых и жанровых характеристик для описания стилевых особенностей коллекций текстов // Новые информационные технологии в автоматизированных системах. 2014. №17.

URL: <http://cyberleninka.ru/article/n/opyt-primeneniya-stilevyh-i-zhanrovyyh-harakteristik-dlya-opisaniya-stilevyh-osobennostey-kollektsiy-tekstov> (дата последнего обращения: 17.05.2017).

46. Прохоров Г. С. Что такое «Художественная публицистика»? // Новый филологический вестник. 2012. №3 (22).

URL: <http://cyberleninka.ru/article/n/chto-takoe-hudozhestvennaya-publitsistika> (дата последнего обращения: 17.05.2017).

47. Современная газетная публицистика. Проблемы стиля // Ответственный редактор И.П.Лысакова, К.А.Рогова. – Л., 1987. С.34

48. Справочник по русскому языку. Практическая стилистика. / Розенталь Д. Э. – М.: издательский дом «ОНИКС 21 век»: Мир и образование, 2001. 381 с.
49. Стилистика русского языка / Под ред. Н. М. Шанского. 2–е изд., доработанное. Л., 1989.
50. Теплова И.И. Специфика преподавания курса «Стилистика русского языка» для студентов–переводчиков // Вестник ННГУ. 2011. №6–2. С.664–666.
- URL: <http://cyberleninka.ru/article/n/spetsifika-prepodavaniya-kursa-stilistika-russkogo-yazyka-dlya-studentov-perevodchikov> (дата последнего обращения: 17.05.2017).
51. Функциональные стили и формы речи / ред. проф. О.Б. Сиротинина – Саратов : Издательство Саратовского университета, 1993. 167 с.
52. Хомутова Т. Н., Петров С. Г. Научно–популярный текст: интегральная модель // Вестник ЮУрГУ. Серия: Лингвистика. 2013. №2.
- URL: <http://cyberleninka.ru/article/n/nauchno-populyarnyy-tekst-integralnaya-model> (дата последнего обращения: 17.05.2017).
53. Шмелев Д. Н. Русский язык в его функциональных разновидностях. М., 1977. С. 34.
54. Apte C., Damerau F. J. and Weiss S.M. 1994. Automated learning of decision rules for text categorization. ACM Trans. on Inform. Syst. 12, 3, 233–251.
55. Bagavandas M., Manimannan G. Style Consistency and Authorship Attribution. A Statistical Investigation // Journal of Quantitative Linguistics. 2008. № 15 (1). P. 100–110.
56. Cleuziou G., Poudat C. On the impact of Lexical and Linguistic features in Genre and Domain–Based Text Categorization / Proceedings of the Eighth International Conference on Intelligent Text Processing and Computational Linguistics, February 2007.

57. Fisher, R.A. The Use of Multiple Measurements in Taxonomic Problems // Annals of Eugenics. — 1936 T. 7. P. 179–188.

58. Garson, G. D. Discriminant Function Analysis. Asheboro, NC: Statistical Associates Publishers. — 2012.

URL:

<https://web.archive.org/web/20080312065328/http://www2.chass.ncsu.edu:80/garson/pA765/discrim.htm> (дата последнего обращения: 17.05.2017)

59. Klecka, William R. Discriminant analysis. Quantitative Applications in the Social Sciences Series, No. 19. Thousand Oaks, CA: Sage Publications. — 1980

60. Manning Chr. D., Raghavan Pr., Schütze H. Introduction to Information Retrieval, Cambridge University Press — 2008

URL: <https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf> (дата последнего обращения: 17.05.2017).

61. Ng A. Stanford CS229 Lecture Notes, 2011. 30 p.

URL: <http://cs229.stanford.edu/notes/cs229-notes1.pdf> (дата последнего обращения: 17.05.2017)

62. Sebastiani F. Text Categorization // Text Mining and Its Applications, WIT Press, Southhampton, UK, 2005. pp.109–129

63. StatSoft. Электронный учебник по статистике [электронный ресурс].

URL: <http://statsoft.ru/home/textbook/modules/stdiscan.html> (дата последнего обращения: 17.05.2017)

## ПРИЛОЖЕНИЕ А. Код программы автоматического определения стилистической принадлежности текстов

```
# -*- coding: UTF-8 -*-
import sys
import codecs
import numpy as np
import argparse
import os
import glob

def createParser ():
    parser = argparse.ArgumentParser()
    parser.add_argument ('-f', '--files', type=str, help=u'путь к файлу / папке с
файлами')
    parser.add_argument ('-r', '--res', type=str, help=u'файл с результатами
обработки файлов', default = 'D:/Result.xml')
    parser.add_argument ('-m', '--mean', action='store', type=str, help=u'вывод
средних значений по всем обработанным файлам – ДА/НЕТ')
    return parser

def statistics (current_file):
    file = codecs.open(current_file, 'r', encoding = 'utf-8')
    nouns = 0.0
    verbs = 0.0
    global adj
    adj = 0.0
    global prcl
    prcl = 0.0
    global npro
    npro = 0.0

    global total_lines
    total_lines = 0.0
    sent_length_count = 0.0
    global verb_noun
    verb_noun = 0.0
    noun_noun = 0.0
    global noun_noun_gen
    noun_noun_gen = 0.0

    global dyn_stat
    dyn_stat = 0.0
    verb_adv = 0.0
    grnd_noun = 0.0
    grnd_adv = 0.0
    adj_noun = 0.0

    global sent
    sent = []
    words = []
```



```

lines = file.readlines()
for index, item in enumerate(lines):
    next_index = index + 1

    if item.find('NOUN') != -1:
        nouns += 1

    elif ((item.find('VERB') != -1) or (item.find('INFN') != -1) or (item.find
('GRND') != -1)):
        verbs += 1

    elif item.find('ADJF') != -1:
        adj += 1

    elif item.find('PRCL') != -1:
        prcl += 1

    elif item.find('NPRO') != -1:
        npro += 1

    if next_index < len(lines):
        next_item = lines[next_index]
        if ((item.find('NOUN') != -1) and (next_item.find('NOUN') != -1))
== True:
            noun_noun += 1
            if (next_item.find('gent') != -1):
                noun_noun_gen += 1
            if (((item.find('NOUN') != -1) and (next_item.find('VERB') != -
1)) or ((item.find('VERB') != -1) and (next_item.find('NOUN') != -1))) == True:
                verb_noun += 1

            if (((item.find('ADVB') != -1) and (next_item.find('VERB') != -
1)) or ((item.find('VERB') != -1) and (next_item.find('ADVB') != -1))) == True:
                verb_adv += 1
            if ((item.find('GRND') != -1) and (next_item.find('NOUN') != -1))
== True:
                grnd_noun += 1
                if (((item.find('GRND') != -1) and (next_item.find('ADVB') != -
1)) or ((item.find('ADVB') != -1) and (next_item.find('GRND') != -1))) == True:
                    grnd_adv += 1
                if ((item.find('ADJF') != -1) and (next_item.find('NOUN') != -1))
== True: # подсчёт конструкций ПРИЛ+СУЩ
                    adj_noun += 1

            if ((item.find('sent') == -1) and (item.find('PNCT') == -1)) :
                total_lines += 1
                sent_length_count += 1
                if ((item.find('LATN') == -1) and (item.find('NUMB') == -1) and
(item.find('ROMN') == -1)) :
                    word_stub = item[(item.index('\t') + 1):]
                    word = word_stub[:word_stub.index('\t')]

```

```

        words.append(float(len(word)))
        words_mean.append(float(len(word)))

    elif (item.find('sent') != -1):
        if sent_length_count != 0.0:
            sent.append(sent_length_count)
            sent_mean.append(sent_length_count)
            sent_length_count = 0.0

nouns_mean.append(nouns)
nouns_mean_ratio.append(nouns/total_lines)

verbs_mean.append(verbs)
verbs_mean_ratio.append(verbs/total_lines)

adj_mean.append(adj)
adj_mean_ratio.append(adj/total_lines)

prcl_mean.append(prcl)
prcl_mean_ratio.append(prcl/total_lines)

npro_mean.append(npro)
npro_mean_ratio.append(npro/total_lines)

verb_noun_mean.append(verb_noun)
noun_noun_mean.append(noun_noun)
noun_noun_gen_mean.append(noun_noun_gen)
total_lines_all.append(total_lines)

dyn_stat = ((verb_noun + verb_adv + grnd_noun + grnd_adv)/(noun_noun +
adj_noun))
dyn_stat_mean.append(dyn_stat)

file2.write(u'Обрабатываемый файл: ' + str(current_file) + '\r\n')

file2.write(u'\r\nСреднее число слов в предложении: ' +
str(round(np.asarray(sent).sum()/float(len(sent)), 4)) + '; min: ' + str(min(sent)) + '; max: ' +
str(max(sent)))

file2.write(u'\r\nСреднее число букв в слове: ' +
str(round(np.asarray(words).sum()/float(len(words)), 4)) + '; min: ' + str(min(words)) + '; max: ' +
str(max(words)))

file2.write(u'\r\nЧисло групп \"существительное + существительное\": ' +
str(noun_noun) + u'; из них \"существительное + существительное в род.п.\": ' +
str(noun_noun_gen) + u'\r\nЧисло групп \"глагол + существительное\": ' + str(verb_noun))

file2.write(u'\r\nЗначение параметра соотношения динамичности и
статичности: ' + str(round((dyn_stat), 4)))

file2.write(u'\r\nВсего слов в тексте: ' + str(total_lines))

```

```

file2.write(u'\r\nСуществительных: ' + str(nouns) + u' Доля (относительно
общего числа слов в тексте): ' + str(round((nouns/total_lines), 4)))

file2.write(u'\r\nГлаголов: ' + str(verbs) + u' Доля (относительно общего числа
слов в тексте): ' + str(round((verbs/total_lines), 4)))

file2.write(u'\r\nПрилагательных: ' + str(adj) + u' Доля (относительно общего
числа слов в тексте): ' + str(round((adj/total_lines), 4)))

file2.write(u'\r\nЛичных местоимений: ' + str(npro) + u' Доля (относительно
общего числа слов в тексте): ' + str(round((npro/total_lines), 4)))

file2.write(u'\r\nЧастиц: ' + str(prcl) + u' Доля (относительно общего числа
слов в тексте): ' + str(round((prcl/total_lines), 4)) + '\r\n\r\n')

print (u'\r\nОбработан файл ' + current_file)

style()

file.close()

def style():
    if (((np.asarray(sent).sum()/float(len(sent))) >= 18) and (dyn_stat <= 0.1)):
        print (u'Скорее всего, это деловой текст')
    elif (((npro/total_lines) >= 0.03) and ((prcl/total_lines) >= 0.02)):
        print (u'Скорее всего, это художественный текст')
    elif (((adj/total_lines) >= 0.1) and ((adj/total_lines) <= 0.2)) and
((prcl/total_lines) <= 0.02) and ((dyn_stat) <= 0.32) and (((npro/total_lines) >= 0.0038) and
((npro/total_lines) <= 0.0231)):
        print (u'Скорее всего, это научный текст')
    else:
        print (u'Скорее всего, это публицистический текст')

parser = createParser()
file = parser.parse_args(sys.argv[1:])
file2 = codecs.open(file.res, 'a', encoding = 'utf-8')
file2.seek(0)
file2.truncate()

verbs_mean = []
verbs_mean_ratio = []

nouns_mean = []
nouns_mean_ratio = []

adj_mean = []
adj_mean_ratio = []

prcl_mean = []
prcl_mean_ratio = []

```

```

npro_mean = []
npro_mean_ratio = []

verb_noun_mean = []
noun_noun_mean = []
noun_noun_gen_mean = []
dyn_stat_mean = []

sent_mean = []
words_mean = []
total_lines_all = []

if ((str(file.files).find('.xml') != -1) and (os.path.exists(file.files) == True)):
    statistics(file.files)
elif ((file.files != None) and (os.path.exists(file.files) == True)) :
    all_files = glob.glob(file.files + '/*.xml')
    for ffile in all_files:
        statistics(ffile)
else:
    print (u'Пожалуйста, введите путь к файлу / папке с файлами')
    sys.exit()

if ((file.mean == 'y') or (file.mean == 'Y')):
    file2.write(u'\r\nСредняя доля существительных (по всем файлам): ' +
str(round(np.asarray(nouns_mean).sum()/float(sum(total_lines_all)), 4)) + '; min: ' +
str(round(min(nouns_mean_ratio), 4)) + '; max: ' + str(round(max(nouns_mean_ratio), 4)) + u';
стандартное отклонение: ' + str(round(np.std(nouns_mean_ratio), 4)) + '\r\n')

    file2.write(u'\r\nСредняя доля глаголов (по всем файлам): ' +
str(round(np.asarray(verbs_mean).sum()/float(sum(total_lines_all)), 4)) + '; min: ' +
str(round(min(verbs_mean_ratio), 4)) + '; max: ' + str(round(max(verbs_mean_ratio), 4)) + u';
стандартное отклонение: ' + str(round(np.std(verbs_mean_ratio), 4)) + '\r\n')

    file2.write(u'\r\nСредняя доля прилагательных (по всем файлам): ' +
str(round(np.asarray(adj_mean).sum()/float(sum(total_lines_all)), 4)) + '; min: ' +
str(round(min(adj_mean_ratio), 4)) + '; max: ' + str(round(max(adj_mean_ratio), 4)) + u';
стандартное отклонение: ' + str(round(np.std(adj_mean_ratio), 4)) + '\r\n')

    file2.write(u'\r\nСредняя доля частиц (по всем файлам): ' +
str(round(np.asarray(prcl_mean).sum()/float(sum(total_lines_all)), 4)) + '; min: ' +
str(round(min(prcl_mean_ratio), 4)) + '; max: ' + str(round(max(prcl_mean_ratio), 4)) + u';
стандартное отклонение: ' + str(round(np.std(prcl_mean_ratio), 4)) + '\r\n')

    file2.write(u'\r\nСредняя доля личных местоимений (по всем файлам): ' +
str(round(np.asarray(npro_mean).sum()/float(sum(total_lines_all)), 4)) + '; min: ' +
str(round(min(npro_mean_ratio), 4)) + '; max: ' + str(round(max(npro_mean_ratio), 4)) + u';
стандартное отклонение: ' + str(round(np.std(npro_mean_ratio), 4)) + '\r\n')

    file2.write(u'\r\nСреднее число конструкций \'гл+сущ\' (по всем файлам): ' +
str(round(np.asarray(verb_noun_mean).sum()/float(len(verb_noun_mean)), 4)) + '; min: ' +

```

```
str(min(verb_noun_mean)) + '; max: ' + str(max(verb_noun_mean)) + u'; стандартное  
отклонение: ' + str(round(np.std(verb_noun_mean), 4)) + '\r\n')
```

```
file2.write(u'\r\nСреднее число конструкций \'сущ+сущ\' (по всем файлам): ' +  
str(round(np.asarray(noun_noun_mean).sum()/float(len(noun_noun_mean)), 4)) + '; min: ' +  
str(min(noun_noun_mean)) + '; max: ' + str(max(noun_noun_mean)) + u'; стандартное  
отклонение: ' + str(round(np.std(noun_noun_mean), 4)) + '\r\n')
```

```
file2.write(u'\r\nСреднее число конструкций \'сущ+сущ в род.п.\' (по всем  
файлам): ' +  
str(round(np.asarray(noun_noun_gen_mean).sum()/float(len(noun_noun_gen_mean)), 4)) + ';  
min: ' + str(min(noun_noun_gen_mean)) + '; max: ' + str(max(noun_noun_gen_mean)) + u';  
стандартное отклонение: ' + str(round(np.std(noun_noun_gen_mean), 4)) + '\r\n')
```

```
file2.write(u'\r\nСреднее значение параметра соотношения динамичности и  
статичности (по всем файлам): ' +  
str(round(np.asarray(dyn_stat_mean).sum()/float(len(dyn_stat_mean)), 4)) + '; min: ' +  
str(round(min(dyn_stat_mean), 4)) + '; max: ' + str(round(max(dyn_stat_mean), 4)) + u';  
стандартное отклонение: ' + str(round(np.std(dyn_stat_mean), 4)) + '\r\n')
```

```
file2.write(u'\r\nСредняя длина предложений (по всем файлам): ' +  
str(round(np.asarray(sent_mean).sum()/float(len(sent_mean)), 4)) + '; min: ' +  
str(min(sent_mean)) + '; max: ' + str(max(sent_mean)) + u'; стандартное отклонение: ' +  
str(round(np.std(sent_mean), 4)) + '\r\n')
```

```
file2.write(u'\r\nСредняя длина слов (по всем файлам): ' +  
str(round(np.asarray(words_mean).sum()/float(len(words_mean)), 4)) + '; min: ' +  
str(min(words_mean)) + '; max: ' + str(max(words_mean)) + u'; стандартное отклонение: ' +  
str(round(np.std(words_mean), 4)) + '\r\n')
```

```
file2.close()  
print (u'\r\nВсе файлы проанализированы.')
```

## **ПРИЛОЖЕНИЕ Б. Перечень текстов, использованных при создании корпусов**

### **Тексты художественного стиля**

1. Нечипоренко В. Агент чужой планеты Ветер [Электронный ресурс] // Библиотека Максима Мошкова [Офиц. сайт]. URL: <http://lib-rus.ru/RUFANT/NECHIPORENKO/agent.txt.html> (дата последнего обращения: 17.05.2017).
2. Носов Е. Солнечный Ветер [Электронный ресурс] // Библиотека Максима Мошкова [Офиц. сайт]. URL: [http://lib-rus.ru/RUFANT/NOSOW\\_E/nosov1.txt.html](http://lib-rus.ru/RUFANT/NOSOW_E/nosov1.txt.html) (дата последнего обращения: 17.05.2017).
3. Петухов Ю. Ангел Возмездия // Библиотека Максима Мошкова [Офиц. сайт]. URL: <http://lib-rus.ru/RUFANT/PETUHOW/revenge.txt.html> (дата последнего обращения: 17.05.2017).
4. Плонский А. По ту сторону вселенной [Электронный ресурс] // Библиотека Максима Мошкова [Офиц. сайт]. URL: [http://lib-rus.ru/RUFANT/PLONSKIJ/po\\_tu.txt.html](http://lib-rus.ru/RUFANT/PLONSKIJ/po_tu.txt.html) (дата последнего обращения: 17.05.2017).
5. Плонский А. Взять высоту! [Электронный ресурс] // Библиотека Максима Мошкова [Офиц. сайт]. URL: <http://lib-rus.ru/RUFANT/PLONSKIJ/vzyat.txt.html> (дата последнего обращения: 17.05.2017).
6. Пронин О. Тайна чёрной планеты [Электронный ресурс] // Библиотека Максима Мошкова [Офиц. сайт]. URL: [http://lib-rus.ru/RUFANT/O\\_PRONIN/p2.txt.html](http://lib-rus.ru/RUFANT/O_PRONIN/p2.txt.html) (дата последнего обращения: 17.05.2017).
7. Полещук А. Ошибка инженера Алексеева [Электронный ресурс] // Библиотека Максима Мошкова [Офиц. сайт]. URL: <http://lib-rus.ru/RUFANT/POLESHUK/error.txt.html> (дата последнего обращения: 17.05.2017).

8. Поляшенко Д. Разведение роз вдали от цивилизации [Электронный ресурс] // Библиотека Максима Мошкова [Официальный сайт]. URL: [http://lib-rus.ru/RUFANT/POLYASHENKO\\_D/soldaty.txt.html](http://lib-rus.ru/RUFANT/POLYASHENKO_D/soldaty.txt.html) (дата последнего обращения: 17.05.2017).

9. Рыбаков В. Гравилет «Цесаревич» [Электронный ресурс] // Библиотека Максима Мошкова [Официальный сайт]. URL: <http://lib.ru/RYBAKOW/gravilet.txt> (дата последнего обращения: 17.05.2017).

10. Скобелев Э. Властелин времени [Электронный ресурс] // Библиотека Максима Мошкова [Официальный сайт]. URL: [http://lib.ru/RUFANT/SKOBELEW\\_E/vlvrem.txt](http://lib.ru/RUFANT/SKOBELEW_E/vlvrem.txt) (дата последнего обращения: 17.05.2017).

11. Скобелев Э. Катастрофа [Электронный ресурс] // Библиотека Максима Мошкова [Официальный сайт]. URL: [http://lib.ru/RUFANT/SKOBELEW\\_E/katastrofa.txt](http://lib.ru/RUFANT/SKOBELEW_E/katastrofa.txt) (дата последнего обращения: 17.05.2017).

12. Снегов С. Галактическая разведка [Электронный ресурс] // Библиотека Максима Мошкова [Официальный сайт]. URL: [http://lib.ru/RUFANT/SNEGOW/asgods\\_1.txt](http://lib.ru/RUFANT/SNEGOW/asgods_1.txt) (дата последнего обращения: 17.05.2017).

13. Снегов С. Вторжение в Персей [Электронный ресурс] // Библиотека Максима Мошкова [Официальный сайт]. URL: [http://lib.ru/RUFANT/SNEGOW/asgods\\_2.txt](http://lib.ru/RUFANT/SNEGOW/asgods_2.txt) (дата последнего обращения: 17.05.2017).

14. Стругацкий А., Стругацкий Б. Путь на Амальтею [Электронный ресурс] // Библиотека Максима Мошкова [Официальный сайт]. URL: <http://lib.ru/STRUGACKIE/amalxteq.txt> (дата последнего обращения: 17.05.2017).

15. Стругацкий А., Стругацкий Б. Частные предположения [Электронный ресурс] // Библиотека Максима Мошкова [Официальный сайт]. URL:

<http://lib.ru/STRUGACKIE/chastnye.txt> (дата последнего обращения: 17.05.2017).

16. Тимофеев П. Обратный отсчет [Электронный ресурс] // Библиотека Максима Мошкова [Офиц. сайт]. URL: [http://fan.lib.ru/t/timofeew\\_p\\_g/countdown.shtml](http://fan.lib.ru/t/timofeew_p_g/countdown.shtml) (дата последнего обращения: 17.05.2017).

17. Томан Н.В. Девушка с планеты Эффа [Электронный ресурс] // Библиотека Максима Мошкова [Офиц. сайт]. URL: [http://lib.ru/RUFANT/TOMAN/govorit\\_cosmos.txt](http://lib.ru/RUFANT/TOMAN/govorit_cosmos.txt) (дата последнего обращения: 17.05.2017).

18. Тюрин А. Меч космонавта, или Сказ об украденном времени [Электронный ресурс] // Библиотека Максима Мошкова [Офиц. сайт]. URL: [http://fan.lib.ru/t/tjurin\\_a\\_w/tyurin-meko.shtml](http://fan.lib.ru/t/tjurin_a_w/tyurin-meko.shtml) (дата последнего обращения: 17.05.2017).

19. Шалимов А. Все началось с "Евы" [Электронный ресурс] // Библиотека Максима Мошкова [Офиц. сайт]. URL: <http://lib.ru/RUFANT/SHALIMOW/allbegin.txt> (дата последнего обращения: 17.05.2017).

20. Шалимов А. Концентратор гравитации [Электронный ресурс] // Библиотека Максима Мошкова [Офиц. сайт]. URL: <http://lib.ru/RUFANT/SHALIMOW/concentr.txt> (дата обращения – 26.05.2015).

21. Шах Г. Гибель Фазтона [Электронный ресурс] // Библиотека Максима Мошкова [Офиц. сайт]. URL: <http://lib.ru/RUFANT/SHAH/53-06.txt> (дата обращения – 26.05.2015).

22. Шах Г. О, марсиане! [Электронный ресурс] // Библиотека Максима Мошкова [Офиц. сайт]. URL: <http://lib.ru/RUFANT/SHAH/53-08.txt> (дата обращения – 26.05.2015).

23. Щепетнёв В. Наш человек на Марсе [Электронный ресурс] // Библиотека Максима Мошкова [Офиц. сайт]. URL:



[http://fan.lib.ru/s/shepetnew\\_wasilij\\_pawlowich/text\\_0010.shtml](http://fan.lib.ru/s/shepetnew_wasilij_pawlowich/text_0010.shtml) (дата обращения – 26.05.2015).

24. Юрин Д. Взлет–Посадка [Электронный ресурс] // Библиотека Максима Мошкова [Офиц. сайт]. URL: [http://fan.lib.ru/j/jurin\\_d/text\\_0010.shtml](http://fan.lib.ru/j/jurin_d/text_0010.shtml) (дата обращения – 26.05.2015).

25. Якубовский А. В складке времени [Электронный ресурс] // Библиотека Максима Мошкова [Офиц. сайт]. URL: <http://lib.ru/RUFANT/YAKUBOWSKIJ/24-06.txt> (дата обращения – 26.05.2015).

### **Тексты научного стиля**

1. Авдуевский В.С., Успенский Г.Р. Космическая индустрия. М.: Машиностроение, 2-е изд., перераб. и дор. , 1989, - 568 с.

2. Алешков М.Н., Жуков И.И., Савин Н.В. и др. Физические основы ракетного оружия. М.: Воениздат, 1972. — 312 с.

3. Аксенов Е.П. Теория движения искусственных спутников земли. М.: Наука. Гл. ред. физ.-мат. лит., 1977, 360 стр.

4. Ануреев И.И. Ракеты многократного использования. — М.: Воениздат, 1975 — 214 с.

5. Балабух Л.И., Алфутов Н.А., Усюкин В.И. Строительная механика ракет: Учебник для машиностроительных спец. Вузов. — М.: Высш. Шк., 1984. — 391 с.

6. Белецкий В.В. Движение искусственного спутника относительно центра масс / Белецкий В. В. – М.: Книга по Требованию, 2012. – 414 с.

7. Беляков И.Т., Борисов Ю.Д. Основы космической технологии. Учеб. Пособие для вузов. — М.: Машиностроение, 1980. — 184 с., ил.

8. Беляков И.Т., Зернов И.А., Антонов Е.Г. и др. Технология сборки и испытаний космических аппаратов – М.: Машиностроение, 1990. – 352с.

9. Гардымов Г.П., Парфенов Б.А., Пчелинцев А.В. Технология ракетостроения. Учебное пособие. – СПб: "Специальная литература", 1997. – 320 с.

10. Гильзин К.А. Электрические межпланетные корабли. — Изд. 2-е, перераб. и доп. — М.: Наука, 1970 — 432 с.
11. Гущин В.Н. Основы устройства космических аппаратов: Учебник для вузов. — М.: Машиностроение, 2003. — 272 с.
12. Дятлов А.П. Системы спутниковой связи с подвижными объектами. Учебное пособие. Ч.1. — Таганрог, ТРТУ, 1997. — 95 с.
13. Иванов Н.М. Баллистика и навигация космических аппаратов. Гриф МО РФ. — М.: Дрофа, 2008.
14. Камалов В.С. Производство космических аппаратов — Москва. Машиностроение 1982. — 280 стр.
15. Куркоткин В.И., Стерлингов В.Л. Самонаведение ракет — М.: Военное издательство Министерства обороны СССР, 1963 — 89 с.
16. Левантовский В.И. Небесная баллистика — М.: Знание, 1965. — 160 с.
17. Саврасов Ю.С. Методы определения орбит космических объектов — М.: Машиностроение, 1981. — 174 с.
18. Семенов А.А. Спускаемая капсула космического аппарата. — СПб.: Нева, 2009.
19. Серебряков В.Н. Основы проектирования систем жизнеобеспечения экипажа космических летательных аппаратов. — М.: Машиностроение, 1983. — 160 с.
20. Святодух В.К. Динамика пространственного движения управляемых ракет - Москва : Машиностроение, 1969. — 273 с.
21. Солодов А.В. Инженерный справочник по космической технике. Москва, ВОЕННОЕ ИЗДАТЕЛЬСТВО МО СССР, 1977 г., 430с.
22. Теория ракетных двигателей: Учебник для студентов машиностроительных специальностей вузов / Алемасов В.Е., Дрегаллин А.Ф., Тишин А.П.; Под ред. Глушко В.П. — М.: Машиностроение, 1980. — 533 с.
23. Усолкин Ю.Ю. Проектирование головных частей баллистических ракет. Учебное пособие - Челябинск: Изд. ЮУрГУ, 2005. — 41 с.

24. Фролов В.С. Инерциальное управление ракетами. – М.: Воениздат, 1975, – 168 с.

25. Энергетика ракетных двигателей на твёрдом топливе / Милёхин Ю.М., Ключников А.Н., Бурский Г.В., Лавров Г.С. — М.: Наука, 2013. — 207 с.

### **Тексты делового стиля**

1. Государственный образовательный стандарт высшего профессионального образования. Направление подготовки дипломированного специалиста 652600 - Ракетостроение и космонавтика. Классификация - инженер.

2. Государственный образовательный стандарт высшего профессионального образования. Направление подготовки дипломированного специалиста 652100 - Авиастроение. Классификация - инженер.

3. Государственный образовательный стандарт высшего профессионального образования. Направление подготовки дипломированного специалиста 211000 - Конструирование и технология электронных средств

4. Государственный образовательный стандарт высшего профессионального образования. Направление подготовки дипломированного специалиста 160700 - Двигатели летательных аппаратов

5. Государственный образовательный стандарт высшего профессионального образования. Направление подготовки дипломированного специалиста 161100 - Системы управления движением и навигация

6. Государственный образовательный стандарт высшего профессионального образования. Направление подготовки дипломированного специалиста 160700 - Двигатели летательных аппаратов

7. Государственный образовательный стандарт высшего профессионального образования. Направление подготовки дипломированного специалиста 210100 - Электроника и нанoeлектроника

8. Государственный образовательный стандарт высшего профессионального образования. Направление подготовки дипломированного специалиста 162700 - Эксплуатация аэропортов и обеспечение полётов воздушных судов

9. Государственный образовательный стандарт высшего профессионального образования. Направление подготовки дипломированного специалиста 161002 – Летная эксплуатация и применение авиационных комплексов

10. Рабочая программа дисциплины (000000601) Маркетинг космической деятельности (Московский авиационный институт (национальный исследовательский университет))

11. Рабочая программа дисциплины (000000165) Физика (Московский авиационный институт (Московский авиационный институт (национальный исследовательский университет))

12. Рабочая программа дисциплины (000003896) Инженерная графика (Московский авиационный институт (национальный исследовательский университет))

13. Рабочая программа дисциплины (000005640) Сопротивление материалов (Московский авиационный институт (национальный исследовательский университет))

### **Тексты публицистического стиля**

1. Аэрокосмическая техника : [журнал] / Мир, №1. 1988
2. Аэрокосмическая техника : [журнал] / Мир, №2. 1988
3. Аэрокосмическая техника : [журнал] / Мир, №3. 1988
4. Аэрокосмическая техника : [журнал] / Мир, №4. 1988
5. Аэрокосмическая техника : [журнал] / Мир, №5. 1988

6. Аэрокосмическая техника : [журнал] / Мир, №6. 1988
7. Аэрокосмическая техника : [журнал] / Мир, №7. 1988
8. Аэрокосмическая техника : [журнал] / Мир, №8. 1988
9. Аэрокосмическая техника : [журнал] / Мир, №9. 1988
10. Аэрокосмическая техника : [журнал] / Мир, №10. 1989
11. Аэрокосмическая техника : [журнал] / Мир, №11. 1989
12. Аэрокосмическая техника : [журнал] / Мир, №12. 1989
13. Аэрокосмическая техника : [журнал] / Мир, №1. 1989
14. Аэрокосмическая техника : [журнал] / Мир, №2. 1989
15. Аэрокосмическая техника : [журнал] / Мир, №3. 1989
16. Аэрокосмическая техника : [журнал] / Мир, №4. 1989
17. Аэрокосмическая техника : [журнал] / Мир, №5. 1989
18. Аэрокосмическая техника : [журнал] / Мир, №6. 1989
19. Аэрокосмическая техника : [журнал] / Мир, №7. 1989
20. Аэрокосмическая техника : [журнал] / Мир, №8. 1989
21. Аэрокосмическая техника : [журнал] / Мир, №9. 1989
22. Аэрокосмическая техника : [журнал] / Мир, №10. 1989
23. Аэрокосмическая техника : [журнал] / Мир, №11. 1989
24. Аэрокосмическая техника : [журнал] / Мир, №12. 1989
25. Новости космонавтики : [журнал]. / Информационно-издательский дом "Новости Космонавтики", №1. 1998
26. Новости космонавтики : [журнал]. / Информационно-издательский дом "Новости Космонавтики", №3. 1998
27. Новости космонавтики : [журнал]. / Информационно-издательский дом "Новости Космонавтики", №6. 1998
28. Новости космонавтики : [журнал]. / Информационно-издательский дом "Новости Космонавтики", №8. 1998
29. Новости космонавтики : [журнал]. / Информационно-издательский дом "Новости Космонавтики", №10. 1998

30.       Новости космонавтики : [журнал]. / Информационно-издательский дом "Новости Космонавтики", №13. 1998