Автоматическое определение стилистической принадлежности текстов по их статистическим параметрам

А.Р. Дубовик

Санкт-Петербургский государственный университет

anna.dubovik0501@gmail.com

Аннотация

Исследование основывается на предположении о том, что принадлежность текста к тому или иному функциональному стилю возможно определить, опираясь на его статистические параметры. Определение стилей осуществляется на основании методов сплошного анализа материалов; разработанная программная реализация позволяет оценить используемые параметры. Эксперименты проводились на материале русскоязычных корпусов текстов, принадлежащих четырём функциональным стилям (научному, художественному, официально-деловому и публицистическому); корпусы были собраны и морфологически аннотированы автором.

Ключевые слова: стиль, классификация, корпусная лингвистика, стилеметрия, типология.

1. 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 [1].

В данной работе мы обратимся к проблеме автоматической классификации текстов на русском языке. Наше исследование основывается на предположении о том, что принадлежность текста к тому или иному функциональному стилю возможно определить, опираясь на статистические данные.

Мы проведём анализ текстов четырёх функциональных стилей, выделим статистические параметры, характеризующие каждый из данных стилей, и создадим собственный компьютерный инструмент, способный проводить автоматическую стилистическую диагностику текстов.

2. Теоретические основания для автоматической стилистической диагностики текстов на русском языке

В нашей работе рассматриваются тексты четырёх стилей: научного, официально-делового, художественного и публицистического. При описании характеристик этих стилей мы будем опираться на работу [2].

2.1. 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, подчёркнутая 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:

- терминологичность, господство обобщённо-отвлечённой лексики;
- специфические параметры распределения частей речи (наименьшая среди всех стилей частотность предикативных форм глагола; общее превалирование имён; преобладание глаголов в форме настоящего времени);
- специфические параметры распределения синтаксических структур (частотность сложных предложений, причастных оборотов, безличных предложений; преобладающее использование комбинированных словосочетаний).

2.2. 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:

- лексическое богатство;
- употребление преимущественно семантически конкретных существительных;
- обилие глаголов говорения, разнообразных частиц и местоимений-существительных;
- высокая частотность форм именительного и родительного падежей имён существительных;

- преимущественное использование простых словосочетаний.

2.3. Деловой стиль

Деловой стиль обслуживает административно-правовую сферу деятельности. Он 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.

Для этого стиля характерны:

- стандартное расположение материала;
- широкое использование терминологии (деловой, юридической), а также официальной, канцелярской лексики и фразеологии, включение в текст сложносокращённых слов, аббревиатур;
- частое употребление отглагольных существительных, производных предлогов (*на основании, в отношении, в соответствии с, за счёт* 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 …; то обстоятельство, что …; тот факт, что …* и т. п.);
- использование номинативных предложений с перечислением;
- использование цепочек родительных падежей (см. пример в работе [3]: *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*);
- тенденция к употреблению сложноподчинённых предложений, отражающих логическое подчинение одних фактов другим;
- почти полное отсутствие эмоционально-экспрессивных речевых средств.

2.4. Публицистический стиль

Публицистический стиль обслуживает сферу политико-идеологических социальных отношений, соотносится с познавательно-оценивающей работой сознания. Его цель — привлечь внимание адресата. Поэтому средства достижения выразительности и экспрессивности в публицистическом стиле имеют большее значение, чем в других стилях.

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.

Для этого стиля характерны:

- наличие общественно-политической лексики и фразеологии, переосмысление лексики других стилей (в частности, терминологической) для целей публицистики;
- использование характерных для данного стиля клише (различные стёртые метафоры, речевые штампы, устойчивые эпитеты (*по данным социологического опроса, жест доброй воли, мрачные прогнозы, горячая поддержка* и пр.));
- использование изобразительно-выразительных средств языка, в частности средств стилистического синтаксиса (риторические вопросы и восклицания, параллелизм, повторы, инверсия).

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.

2.5. Подбор характеризующих признаков

Для проведения экспериментов мы приняли за основу индексы, предложенные А.Ф. Журавлевым в работе «Опыт квантитативно-типологического исследования разновидностей устной речи» [4], и индексы, предложенные М.А. Марусенко в работе «Атрибуция анонимных и псевдонимных литературных произведений методами теории распознавания образов» [5].

Мы реализовали подсчёт следующих параметров для текстов исследуемых корпусов:

- глагольность: отношение числа глаголов к числу слов в тексте;
- субстантивность: отношение числа существительных к числу словоформ в тексте;
- адъективность: отношение числа прилагательных к числу словоформ в тексте;
- отношение числа личных местоимений к числу словоформ в тексте;
- отношение числа частиц к числу словоформ в тексте;
- отношение числа междометий к числу словоформ в тексте;
- количество конструкций «существительное + существительное» (в том числе количество конструкций «существительное + существительное в родительном падеже»);
- количество конструкций «глагол + существительное»;
- комбинированный параметр, отражающий соотношение динамичности/статичности текстов коллекции, предложенный в работе [6];
- средняя длина слова (число символов от пробела до пробела);
- средняя длина предложения (от точки до точки).

3. Компьютерный инструмент для проведения статистической обработки русскоязычных текстов

3.1. Используемое программное обеспечение

Разработанная нами программа написана на языке программирования Python, что позволяет ей работать практически на всех платформах. Для корректной работы программы необходим Python версии 2.7 и установленная библиотека NumPy. 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 [7].

3.2. Требования к входным данным

На вход программе подаётся результат работы гибридного морфоанализатора NLTK4RUSSIAN: файл с расширением .xml, в котором каждая строка — либо тег начала/конца предложения, либо пунктуационный знак, либо словоформа с аннотацией. Данный файл должен быть сохранён в кодировке UTF-8.

3.3. Ход экспериментов

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

- научный стиль — тексты по радиоэлектронике, ракетостроению и технике;
- художественный стиль — научно-фантастические произведения второй половины XX – начала XXI века;
- 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.;
- публицистический стиль — статьи из выпусков журналов «Новости космонавтики» и «Аэрокосмическая техника» за 1989-2000 гг.

Объём каждого корпуса — 500 тыс. словоупотреблений.

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 квантитативные 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 полученных 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.

Таблица 1. Распределение имён существительных по подкорпусам

Параметр	Художественный стиль	Научный стиль	Публицистический стиль	Деловой стиль
min	0,2264	0,3394	0,2949	0,4165
max	0,3267	0,4552	0,3859	0,5152
mean	0,2763	0,392	0,3583	0,4504
StD	0,2570	0,0238	0,0182	0,0193

Таблица 2. Распределение имён прилагательных по подкорпусам

Параметр	Художественный стиль	Научный стиль	Публицистический стиль	Деловой стиль
min	0,0709	0,1152	0,0941	0,1400
max	0,1436	0,1980	0,1474	0,2170
mean	0,1155	0,1512	0,1217	0,1960
StD	0,0155	0,0195	0,0135	0,0139

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 (например, *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.).

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 (например, *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.

Таблица 3. Распределение глаголов по подкорпусам

Параметр	Художественный стиль	Научный стиль	Публицистический стиль	Деловой стиль
min	0,1473	0,0493	0,0779	0,0208
max	0,1983	0,1013	0,1252	0,0613
mean	0,1670	0,0791	0,0962	0,0505
StD	0,0138	0,0135	0,0129	0,0086

Согласно таблице 3, 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
min	0,0446	0,0038	0,0087	0,0024
max	0,1117	0,0231	0,0446	0,0113
mean	0,0695	0,0124	0,0183	0,0065
StD	0,0159	0,0040	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
min	0,0232	0,0032	0,0089	0,0001
max	0,0673	0,0284	0,0278	0,0064
mean	0,0491	0,0118	0,0169	0,0026
StD	0,0105	0,0051	0,0050	0,0017

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 (например, 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 [6] и [8]. 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 лексемы (подробнее о контекстных профилях см. [9] [10] [11]).

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.

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

Параметр	Художественный стиль	Научный стиль	Публицистический стиль	Деловой стиль
min	193	690	552	941
max	498	1304	1035	1496
mean	349,7	1038,15	845,59	1179,26
StD	89,05	150,22	91,76	144,14

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

Параметр	Художественный стиль	Научный стиль	Публицистический стиль	Деловой стиль
min	107	535	350	816
max	357	1105	705	1393
mean	229,62	863,41	587,15	1015
StD	69,67	145,87	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. Высокие значения этого параметра для текстов научного стиля объясняются тем, что конструкция «существительное + существительное в родительном падеже» — это генитивная конструкция, часто соответствующая неоднословному термину (например, *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.

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.

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

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

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 [6] был определён комбинированный параметр β, 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. 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 ($0{,}49 \in [0{,}48; 1{,}26]$).

Таблица 9. Распределение значений параметра β по подкорпусам

Параметр	Художественный стиль	Научный стиль	Публицистический стиль	Деловой стиль
min	0,4813	0,1300	0,2451	0,0038
max	1,2575	0,3195	0,4963	0,0687
mean	0,7844	0,2160	0,3279	0,0388
StD	0,1712	0,0546	0,0679	0,0157

Полученные данные свидетельствуют о возможности использования данного параметра для определения стиля исследуемого текста. К такому же выводу приходят и авторы работы [6].

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.

Таблица 10. Распределение длин слов по подкорпусам

Параметр	Художественный стиль	Научный стиль	Публицистический стиль	Деловой стиль
min	1	1	1	1
max	40	34	37	39
mean	5,64	6,75	6,33	7,99
StD	3,24	3,85	3,73	4,50

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 меры (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 подкорпусам. 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
min	1	1	1	1
max	95	174	262	3130
mean	9,39	16,62	16,27	44,6
StD	7,28	10,83	11,50	139,48

Наибольшее значение данного параметра наблюдается в предложениях делового стиля. Отличительными особенностями текстов данного стиля 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 характеризующих параметров текстов различных функциональных стилей речи нами был разработан программный модуль, позволяющий провести автоматическую диагностику исследуемого текста. За основу был взят алгоритм деревьев принятия решений. Предполагается, что, последовательно сравнивания статистические параметры текста с пороговыми значениями, полученными после анализа результатов экспериментов, можно пройти по дереву и определить стиль обрабатываемого текста.

Результаты работы с различными параметрами 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.

Алгоритм работы модуля определения стилистической принадлежности текста представлен на рис.1.

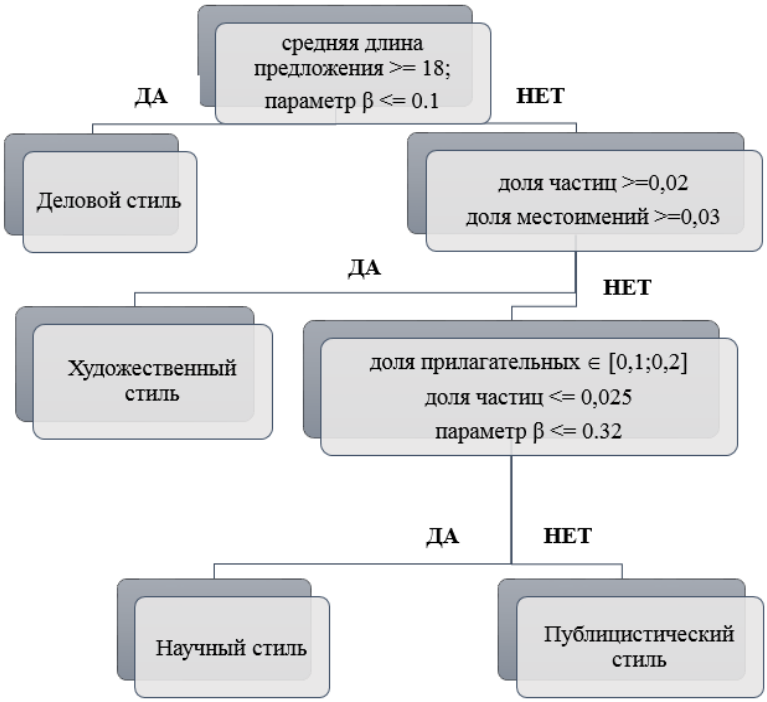

Рис. 1. Алгоритм работы модуля определения стилистической принадлежности текста

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).

3.6. 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

Стиль текста	Кол-во текстов, для которых стиль определён верно	Кол-во текстов, для которых стиль определён неверно
Деловой	65	0
Художественный	65	0
Научный	62	3
Публицистический	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. 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.

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 β 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

Стиль текста	Точность	Полнота	F–мера
Деловой	1	1	1
Художественный	0,99	1	0,99
Научный	0,73	0,95	0,83
Публицистический	0,93	0,57	0,70

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, использовав 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.

3. Заключение

В данной работе мы подробно описали характерные особенности четырёх стилей русского языка — научного, официально-делового, художественного и публицистического — и выдвинули гипотезу 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, во-первых, 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. Во-вторых, можно расширить экспериментальный материал и провести исследования по автоматической обработке бóльшего числа корпусов текстов из других коллекций (например, текстов разговорного стиля или текстов, относящихся к различным литературным жанрам).

Литература

[1] Sebastiani F. Machine learning in automated text categorization // ACM Computing Surveys, 34(1):1–47, 2002 URL: http://nmis.isti.cnr.it/sebastiani/Publications/ACMCS02.pdf.

[2] Функциональные стили и формы речи / ред. проф. О.Б. Сиротинина. – Саратов : Издательство Саратовского университета, 1993. 167 с.

[3] 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.

[4] Журавлев А.Ф. Опыт квантитативно–типологического исследования разновидностей устной речи // Разновидности городской устной речи: Сборник научных трудов. М.: Наука, 1988. С. 84–150.

[5] Марусенко М.А. Атрибуция анонимных и псевдонимных литературных произведений методами распознавания образов. Л.: Изд–во Ленингр. ун–та, 1990. 164 с.

[6] 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. С. 80–85. URL: http://webground.su/data/lit/antonova_klyshinsky_yagunova/Opredeleniye_stilevyh_i_zhanrovyh_kharakteristik.pdf .

[7] 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.

[8] 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.

[9] Веревкина 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. С. 33–39.
[10] 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. С. 370–382.
[11] 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.

Automatic text style identification in terms of statistical parameters

A. Dubovik

Saint-Petersburg State University

This survey presents the main approaches to automatic text style identification. A program suggested here allows text style identification based on quantitative parameters of different styles, such as: average word length, average sentence length, ratio of different parts of speech to the length of the text in question. The experiments were carried out on four corpora of Russian texts representing different styles: news style texts, science-fiction texts, scientific texts and official documents. The experiments show that 88% of texts were classified successfully, best results achieved for official documents and science fiction texts.

Keywords: style, classification, corpus linguistics, stylometry.