Некоторые особенности литературных текстов при их сопоставлении для определения их авторства

Г. Н. Ахобадзе, Е. Ю. Русяева

Институт проблем управления им. В. А. Трапезникова РАН, Москва, Россия

Аннотация. Разработан метод анализа литературных авторских текстов на основе выбора наиболее частотных, характерных для определенного авторского стиля служебных частей речи, и вычисления их весовых коэффициентов (базируется на вычислении наиболее часто используемых в литературных произведениях предлогов, союзов и частиц). Сам процесс вычисления весовых коэффициентов, определяемых отношениями величин служебных частей речи в тексте к общему объему слов, подробно проанализирован. Приводятся экспериментальные результаты по установлению авторства литературных текстов для двух авторов. Результаты получены путем сопоставлений числовых значений однотипных весовых коэффициентов, выраженных в процентах. Полученные теоретические и практические результаты могут быть использованы для анализа, выявления лингвистических особенностей, отличий не только художественных текстов, но, в дальнейшем, текстов любого жанра и стиля.

Ключевые слова: весовой коэффициент, служебные слова, авторство, текст, показатель идентичности, повторяемость.

DOI 10.14357/20718632240207 **EDN** WGWOTS

Введение

Современный мир недаром называют «эпохой ChatGPT4», так как с появлением технологий подобного типа[1] можно говорить о новом витке возможностей развития, коллаборации, гибридных технологий естественного (человеческого) интеллекта (ЕИ) и искусственного интеллекта (ИИ, artificial intelligence - AI). Действительно, ныне разработки генеративного искусственного интеллекта (ГИИ, GAI) во всем мире перешагнули на новый уровень, с их помощью можно создавать все больше и больше текстов на естественном языке. Тем актуальнее становится вопрос об атрибуции текста, ведь проблема определения авторства текста не становится менее актуальной, хотя разработано множество методов и специальных инструментов, помогающих отличать текст одного автора от текста другого, в том числе и созданного искусственно.

Исследований по этой теме, как отечественных, так и зарубежных, множество. Нашей целью не была задача подтвердить или опро-

[1] В нашу задачу не входит детализация разработок по созданию текстов, их классификация и изучение. Данное исследование носит частный характер и является как бы прелюдией к целому циклу исследований. Объектом анализа в этом случае будет авторский текст, поэтому все автоматизированные программы по созданию текстов мы будем обобщенно называть технологиями ИИ, ГИИ, ChatGPT.

вергнуть гипотезу, по которой наиболее удобным стилеметрическим инструментом при определении авторства текста могут быть частоты служебных слов в текстах литературно-художественного стиля, на эту тему есть много интересных исследований [1 - 5]. Мы с таким подходом вполне солидарны и именно на нем базируется наш метод. Нам было важно рассмотреть разные вычислительные варианты этого формального анализа текста, чтобы определить те маркеры среди служебных слов, которые помогут отличить стиль письма одного автора от стиля другого наиболее простым способом, чтобы процент ошибки был минимален. Нами был применен метод вычисления весовых коэффициентов, определяемых отношениями величин выявленных маркеров в тексте к его общему объему, и проведен сопоставительный анализ текстов разных авторов. В качестве маркеров нами были взяты такие наиболее часто встречающиеся в тексте простые служебные части речи в целях избегания омонимии, ведь, по признанию исследователей [6, 7], пока проблем с семантическими анализаторами гораздо больше, чем решений. Также для корректности сопоставления весовых коэффициентов этих маркеров, которые мы назвали идентификаторами идентичности, исходные тексты произведений были сокращены, чтобы общее количество слов в текстах было примерно равным. Для того, чтобы доказать правомерность такого подхода, нами были детально проанализированы по три произведения известных отечественных писателей фантастов С. Лукьяненко и М. Глебова. Задача проверки работы метода на текстах с неизвестным авторством пока не ставилась, мы ищем подходы автоматического лингвистического анализа для повышения точности и упрощения самого процесса вычисления, говорим лишь о некоторых выявленных нами особенностях литературных текстов при их сопоставлении для определения их авторства. Это лишь часть разрабатываемой нами вычислительной системы анализа текстов. В дальнейшие планы входит также разработка более полного комплекса частных методов лингвистического анализа текстов.

1. Краткий анализ разработок по вычислительной обработке текстов естественного языка

Целью этого этапа исследования стала разработка подходов к созданию прототипа информационно-вычислительной системы вычислительного анализа литературных текстов с использованием методов формальной обработки. То, что эта тема вызывает огромный интерес и привлекает внимание исследователей во всем мире, не вызывает сомнений. Вычислительных методов лингвистического анализа сейчас множество, и они разнообразны. Что касается истории вопроса подобного рода исследований, то в начале XX столетия впервые учет статистики употребления пар элементов любой природы, идущих друг за другом в тексте (букв, морфем, словоформ и т.п.), проведен Морозовым Н.А. для расчета лингвистических спектров [8]. Далее к развитию темы числовых оценок для поэтического текста подключился известный русский математик Марков А.А. [9]. Затем новый всплеск активности количественного вычисления авторства текста пришелся на начало XXI, когда появились расчеты Дж. Барроуза, получившие название метода «Дельфи» [10, 11]. Эти вычисления авторства текста также проводились по частотным показателям, причем, среди всех слов текста, и знаменательных, и служебных, наиболее частотны служебные слова, именно с их помощью можно отличать тексты одного автора от другого, для чего и стали активно разрабатываться специальные инструменты анализа. В определенном смысле разрабатываемые стилеметрические инструменты при определении авторства текста на основе вычисления частоты служебных слов в текстах литературно-художественного стиля и есть продолжение, модификация прежних методов. Укажем наиболее заметные на сегодняшний день работы [1-3], и некоторые другие [12, 13].

Так, в исследовании [13] проблему определения авторского литературного стиля текста решают математически, используя сравнение случайности двух заданных текстов. Здесь используется подход, основанный на включении известного двухвыборочного теста Фридмана-Рафски в многоэтапную процедуру с целью ста-

билизации процесса. Широко известен также автоматизированный анализ лексического состава текста с применением частотных характеристик [14], когда исследуются частота использования определенных слов или фраз, уникальные слова и другие лексические характеристики. При создании алгоритмов машинного обучения для анализа авторства текстов собираются данные о текстах разных авторов, и на основе этих данных создается модель [14], которая может, в том числе, даже предсказывать авторство новых текстов. Метод анализа структуры и организации текста больше нацелен на синтаксический анализ, учитывающий исследование способа индивидуальной авторской манеры организации смыслового пространства текста, на то, как автор излагает свои мысли и как структурирует свои тексты. Также анализируются способы синтаксического членения текста на абзацы, главы, части или разделы и т. д, методом сегментации [15].

В большой степени используется статистический анализ, например, даже частотный анализ слов или символов, чтобы найти уникальные характеристики авторского письма на больших текстовых выборках [16]. А также интересны совсем недавние исследования по созданию программных решений для создания лингвистического корпуса текстов на основе обработки естественного языка [17], хотя эти работы все же больше направлены на масштабирование выборки для лингвистического анализа текстов. Отметим, что анализу художественных текстов много внимания уделяется именно в отечественных исследованиях, что связано с тем, что по философской познавательной глубине создания смыслов в России и прошлых веков, и современности, главная роль отводилась художественной литературе. Отечественной познавательной традиции всегда был свойственен глубокий литературный анализ, очень известны в этом плане работы М.Л. Гаспарова, стоит отметить, например, работу [18].

Мы отталкивались от статистического анализа частотности встречаемости тех или иных частей речи в литературном произведении разных авторов. В соответствии с этим подходом мы провели выборочный лингвистический анализ на частоту встречаемости различных слов, а именно, знаменательных и служебных частей речи в текстах различных авторов, работающих в жанре современной фантастики. Наш метод по установлению авторства литературных текстов ни в коей мере не претендует на универсальность. Как указано выше, публикаций, посвященных проблеме установки авторства текстов множество, мы ссылаемся лишь на некоторые, близкие нам по тематике, труды отечественных исследователей, а это, в первую очередь, труды [1-3, 15].

Отметим, что мы делали акцент не на количестве анализируемых текстов, и не на вычислении количества наиболее частотных слов в них. Хотя и мы тоже остановились на служебных словах, которые могут характеризовать идиостилевой профиль писателя, мы среди служебных слов выбирали наиболее частотные и те, которые не омонимичны другим частям речи. На наш взгляд именно они могут стать маркерами, использоваться как идентификаторы (при вычислении их весовых коэффициентов), характеризующие некоторые особенности авторского литературного текста при сопоставлении с текстами другого автора.

2. Теоретическая часть

Основными критериями для создания нашей методики лингвистического анализа литературных текстов стали простота, то есть, упрощение процесса вычисления, и повышение точности (без семантических искажений) анализа. Для выявления наиболее показательных маркеров, характеризующих идиостилевой профиль автора, была проведена предварительная предобработка текста. Этот предварительный анализ показал[2], что помимо того, что наиболее частотными среди всех частей речи являются служебные слова, среди самих служебных слов наиболее часто и с наилучшими показателями (не омонимичными с другими частями речи) стали простые, односоставные служебные части речи: среди предлогов - «в», «на», «с», среди союзов - «и», «а», «но», и среди частиц - «не», «бы», «же». Но оставался вопрос об искажениях в вычислении

[2] Этот вычислительный анализ был проведен с помощью ранее созданного программного продукта «Лингвистический анализатор текста» [19], но детальное описание принципа его работы не входит в задачи данного исследования, мы указываем только результат

в силу разного словарного объема авторских текстов, подлежащих анализу. Вот почему далее, для адекватности сравнения, путем фрагментного анализа, мы поделили тексты на крупные смысловые фрагменты с примерно одинаковым количеством слов, а затем применили метод определения весовых коэффициентов частоты (интенсивностей) использования выделенных нами самых частотных служебных слов. Таким методом мы проанализировали тексты с близкими сюжетами двух известных отечественных авторов, работающих в фантастическом жанре. Затем сравнили вычисленные коэффициенты между собой. Эти вычислительные показатели приведены далее в таблицах для примерно равных объемов текстов с подсчетом количества служебных слов, а именно, предлогов «в», «с» и «на», союзов «и», «а» и «но» и частиц «не», «бы» и «же».

После этого приводится анализ отношений: $n_{1в}/N_1 = K_{1в}$, $n_{1на}/N_1 = K_{1на}$ и $n_{1с}/N_1 = K_{1с}$, где $n_{1в}$ – количество предлогов «в» в первом тексте, $n_{1на}$ – количество предлогов «на» в первом тексте, $n_{1с}$ – количество предлогов «с» в первом тексте, N_1 – количество слов в первом тексте. $n_{1и}/N_1 = K_{1и}$, $n_{1а}/N_1 = K_{1а}$ и $n_{1но}/N_1 = K_{1но}$, где $n_{1и}$ – количество союзов «и» в первом тексте, $n_{1а}$ – количество союзов «а» в первом тексте, $n_{1но}$ – количество союзов «но» в первом тексте. $n_{1не}/N_1 = K_{1не}$, $n_{1бы}/N_1 = K_{1бы}$ и $n_{1же}/N_1 = K_{1же}$, где: $n_{1не}$ – количество частиц «не» в первом тексте, $n_{1бы}$ – количество частиц «бы» в первом тексте, $n_{1же}$ – количество частиц «же» в первом тексте.

Далее вычисляются весовые коэффициенты $K_{1в}$; $K_{1на}$; $K_{1с}$; $K_{1и}$; $K_{1а}$; $K_{1но}$ и $K_{1не}$; $K_{1бы}$; $K_{1же}$ выше указанных предлогов, союзов и частиц. С целью корректности анализа авторства анализируемых трех текстов, словесный объем второго и третьего текстов ограничивают количеством N_1, т.е. количество слов второго и третьего текстов должно равняться N_1 (подсчет ведется от начала рассматриваемых текстов).

Здесь с учетом количества слов N_1, аналогичным образом подсчитываем количество предлогов $n_{2в}$, $n_{2на,}$ $n_{2с}$ во втором тексте и предлогов $n_{3в}$, $n_{3на,}$ $n_{3с}$ в третьем тексте; союзов $n_{2и}$, $n_{2а,}$ $n_{2но}$ во втором тексте и союзов $n_{3и}$, $n_{3а,}$ $n_{3но}$ в третьем тексте; частиц $n_{2не}$, $n_{2бы,}$ $n_{2же}$ во втором тексте и частиц $n_{3не}$, $n_{3бы,}$ $n_{3же}$ в третьем тексте. Далее вычисляем весовые коэффициенты предлогов второго и третьего текстов, соответственно, $K_{2в}$; $K_{2на}$; $K_{2с}$, $K_{3в}$; $K_{3на}$; $K_{3с}$; союзов второго и третьего текстов $K_{2и}$; $K_{2а}$; $K_{2но}$, $K_{3не}$; $K_{3бы}$; $K_{3же}$; частиц второго и третьего текстов $K_{2не}$; $K_{2бы}$; $K_{2же}$, $K_{3не}$; $K_{3бы}$; $K_{3же}$ характеризующие второй и третий анализируемые тексты.

Для окончательного анализа идентификаторов авторства рассматриваемых текстов далее производим сравнение вычисленных однотипных весовых коэффициентов. Другими словами, в данном случае сравниваются коэффициенты: $K_{1в}$, $K_{2в}$ и $K_{3в}$; $K_{1на}$, $K_{2на}$ и $K_{3на}$; $K_{1с}$, $K_{2с}$ и $K_{3с}$; $K_{1и}$, $K_{2и}$ и $K_{3и}$; $K_{1а}$, $K_{2а}$ и $K_{3а}$; $K_{1но}$, $K_{2но}$ и $K_{3но}$; $K_{1не}$, $K_{2не}$ и $K_{3не}$; $K_{1бы}$, $K_{2бы}$ и $K_{3бы}$; $K_{1же}$, $K_{2же}$ и $K_{3же}$. При этом при равных значениях, по меньшей мере у указанных выше однотипных весовых коэффициентов, можно констатировать атрибуцию трех текстов на единое авторство.

В ряде случаев для удобства обозначения размерности весовых коэффициентов (как относительных величин), они могут быть выражены в процентах. Для этого числовые значения весовых коэффициентов следует умножить на 100 %.

3. Практическая часть

Для корректности сравнения идиостилевого профиля авторов, во избежание нарушения авторских прав, тексты авторов фантастических произведений брались в форматах текстовых документов (*.txt) и Word (doc.) с официального сайта «Литрес» [20]. Расчет весовых коэффициентов встречаемости указанных служебных частей речи показал (некоторые выборки представлены далее в таблицах), что сравнительный анализ процентного соотношения весовых коэффициентов встречаемости указанных нами маркерами служебных слов служит не только как наиболее удобный стилеметрический инструмент при определении авторства текста, но и может отчасти упростить процесс вычислений и несколько повысить точность анализа.

3.1. Анализ произведений С. Лукьяненко

Для иллюстрации данного подхода были проанализированы три произведения С. Лукьяненко. Сначала мы применили наш метод для анализа трех из недавно созданных (2020-2022 годов) романов Сергея Лукьяненко из серии «Измененные».

1-й текст С. Лукьяненко «Лето волонтера».

Табл.1. Анализ встречаемости в тексте произведения «Лето волонтера» С. Лукьяненко предлогов «в», «на», «с». Статистика: количество слов 65500 (текст сокращен для корректности сопоставления)

№	Словарный объем текста N (число слов в документе, статистика)	Частота появления предлогов			Весовые коэффициенты предлогов (К)		
		в	на	с	$К_в$	$К_{на}$	$К_с$
1	65500	1512	1038	679	0,0230	0,0158	0,0103
	Встречаемость в тексте в %				2,23%	1,58%	1,03%

Табл. 2. Анализ встречаемости в тексте произведения «Лето волонтера» С. Лукьяненко союзов «и», «а», «но». Статистика: количество слов 65500 (текст сокращен)

№	Словарный объем текста N (число слов в документе, статистика)	Частота появления союзов			Весовые коэффициенты союзов (К)		
		и	а	но	$К_и$	$К_а$	$К_{но}$
1	65500	1998	419	566	0,0305	0,0063	0,0086
	Встречаемость в тексте в %				3,05%	0,63%	0,86%

Табл. 3. Анализ встречаемости в тексте произведения «Лето волонтера» С. Лукьяненко частиц «не», «бы», «же». Статистика: количество слов 65500 (текст сокращен)

№	Словарный объем текста N (число слов в документе, статистика)	Частота появления частиц			Весовые коэффициенты частиц (К)		
		не	бы	же	$К_{не}$	$К_{бы}$	$К_{же}$
1	65500	1572	170	164	0,0240	0,0025	0,0025
	Встречаемость в тексте в %				2,4%	0,25%	0,25%

2-й текст С. Лукьяненко «Месяц за рубиконом»

Табл. 4. Анализ встречаемости в тексте произведения «Месяц за рубиконом» С. Лукьяненко предлогов «в», «на», «с». Статистика: количество слов 65500 (текст сокращен)

№	Словарный объем текста N (число слов в документе, статистика)	Частота появления предлогов			Весовые коэффициенты предлогов (К)		
		в	на	с	$К_в$	$К_{на}$	$К_с$
1	65500	1412	1146	716	0,0215	0,0174	0,010
	Встречаемость в тексте в %				2,15%	1,74%	1,01%

Табл. 5. Анализ встречаемости в тексте произведения «Месяц за рубиконом» С. Лукьяненко союзов «и», «а», «но». Статистика: количество слов 65500 (текст сокращен)

№	Словарный объем текста N (число слов в документе, статистика)	Частота появления союзов			Весовые коэффициенты союзов (К)		
		и	а	но	$К_и$	$К_а$	$К_{но}$
1	65500	1960	397	539	0,0299	0,0060	0,0082
	Встречаемость в тексте в %				2,99%	0,60%	0,82%

Табл. 6. Анализ встречаемости в тексте произведения «Месяц за рубиконом» С. Лукьяненко частиц «не», «бы», «же». Статистика: количество слов 65500 (текст сокращен)

№	Словарный объем текста N (число слов в документе, статистика)	Частота появления частиц			Весовые коэффициенты частиц (К)		
		не	бы	же	$К_{не}$	$К_{бы}$	$К_{же}$
1	65500	1498	208	167	0,0228	0,0031	0,0025
	Встречаемость в тексте в %				2,28%	0,31%	0,25%

3-й текст С. Лукьяненко «Три дня Индиго»

Табл. 7. Анализ встречаемости в тексте произведения «Три дня Индиго» С. Лукьяненко предлогов «в», «на», «с». Статистика: количество слов 65500 (текст сокращен)

№	Словарный объем текста N (число слов в документе, статистика)	Частота появления предлогов			Весовые коэффициенты предлогов (K)		
		в	на	с	$K_в$	$K_{на}$	$K_с$
1	65500	1337	1018	646	0,0204	0,0155	0,0098
Встречаемость в тексте в %					2,04%	1,55%	0,98%

Табл. 8. Анализ встречаемости в тексте произведения «Три дня Индиго» С. Лукьяненко союзов «и», «а», «но». Статистика: количество слов 65500 (текст сокращен)

№	Словарный объем текста N (число слов в документе, статистика)	Частота появления союзов			Весовые коэффициенты союзов (K)		
		и	а	но	$K_и$	$K_а$	$K_{но}$
1	65500	1933	476	507	0,0295	0,0072	0,0077
Встречаемость в тексте в %					2,95%	0,72%	0,77%

Табл. 9. Анализ встречаемости в тексте произведения «Три дня Индиго» С. Лукьяненко частиц «не», «бы», «же». Статистика: количество слов 65500 (текст сокращен)

№	Словарный объем текста N (число слов в документе, статистика)	Частота появления частиц			Весовые коэффициенты частиц (K)		
		не	бы	же	$K_{не}$	$K_{бы}$	$K_{же}$
1	65500	1614	251	143	0,0246	0,0038	0,0021
Встречаемость в тексте в %					2,46%	0,38%	0,21%

Исходя из того, что все три выше исследуемые произведения принадлежат одному автору (С. Лукьяненко), было предложено сопоставить однотипные весовые коэффициенты согласно приведенным выше табличным данным с целью выявления равенства между ними или их незначительной разностью. При этом необходимо, чтобы все тексты имели примерно одинаковое количество слов. В итоге выявлено равенство коэффициентов первого произведения («Лето волонтера») и второго произведения («Месяц за Рубиконом») соответственно $k_{1же} = k_{2же}$ (0,25% = 0,25%), первая разность 0,02 % коэффициентов перового и второго произведений соответственно $k_{1с}$ = 1,03% и $k_{2с}$ = 1,01% (0,02%), а также вторая разность первого произведения и третьего произведения («Три дня Индиго»), соответственно, $k_{1на}$ = 1,58 % и $k_{3на}$ = 1,55% (0,03%).

Другими словами, установлен показатель (критерий) идентичности на основе известных трех произведений и этот показатель может меняться от 0 до 0,03 процентов, который далее может быть использован для установления авторства трех произведений другого автора, в данном случае это произведения М. Глебова. Отсюда можно заключить, что повторяемость (частота) показателя идентичности в трех произведениях С. Лукьяненко, изменяющегося от 0 до 0,03 процентов, составляет 3.

3.2. Анализ произведений М. Глебова

Были исследованы особенности текстов трех его произведений, тексты также скачаны в электронном виде с официального сайта «Литрес» [20]. Проанализированы тексты недавно созданных (2022-2023 годы) фантастических романов М. Глебова серии «Блюстители хаоса». Количество слов также 65500, тексты также были сокращены для корректности сопоставления.

1-й текст М. Глебова «Столица мятежной окраины»

Табл. 10. Анализ встречаемости в тексте произведения «Столица мятежной окраины» серии «Блюстители хаоса» М. Глебова предлогов «в», «на», «с». Статистика: количество слов 65500 (текст сокращен)

№	Словарный объем текста N (число слов в документе, статистика)	Частота появления предлогов			Весовые коэффициенты предлогов (K)		
		в	на	с	$K_в$	$K_{на}$	$K_с$
1	65500	1596	891	688	0,0243	0,0135	0,0104
	Встречаемость в тексте в %				2,43%	1,35%	1,04%

Табл. 11. Анализ встречаемости в тексте произведения «Столица мятежной окраины» М. Глебова союзов «и», «а», «но». Статистика: количество слов 65500 (текст сокращен)

№	Словарный объем текста N (число слов в документе, статистика)	Частота появления союзов			Весовые коэффициенты союзов (K)		
		и	а	но	$K_и$	$K_а$	$K_{но}$
1	65500	1829	383	777	0,0279	0,0058	0,0118
	Встречаемость в тексте в %				2,79%	0,58%	1,18%

Табл. 12. Анализ встречаемости в тексте произведения «Столица мятежной окраины» М. Глебова частиц «не», «бы», «же». Статистика: количество слов 65500 (текст сокращен)

№	Словарный объем текста N (число слов в документе, статистика)	Частота появления частиц			Весовые коэффициенты частиц (K)		
		не	бы	же	$K_{не}$	$K_{бы}$	$K_{же}$
1	65500	1739	209	145	0,0265	0,0031	0,0022
	Встречаемость в тексте в %				2,65%	0,31%	0,22%

2-й текст М. Глебова «Эхо орбитального удара»

Табл. 13. Анализ встречаемости в тексте произведения «Эхо орбитального удара» (серии «Блюстители хаоса» М. Глебова предлогов (так называемых «пространственных», указывающих на расстояние) «в», «на», «с». Статистика: количество слов 65500 (текст сокращен)

№	Словарный объем текста N (число слов в документе, статистика)	Частота появления предлогов			Весовые коэффициенты предлогов (K)		
		в	на	с	$K_в$	$K_{на}$	$K_с$
1	65500	1535	1020	721	0,0234	0,0155	0,0110
	Встречаемость в тексте в %				2,34%	1,55%	1,10%

Табл. 14. Анализ встречаемости в тексте произведения «Эхо орбитального удара» М. Глебова (так называемых «соединительных» союзов, встречающихся в словосочетаниях и сочинительных предложениях) - союзов «и», «а», «но». Статистика: количество слов 65500

№	Словарный объем текста N (число слов в документе, статистика)	Частота появления союзов			Весовые коэффициенты союзов (K)		
		и	а	но	$K_и$	$K_а$	$K_{но}$
1	65500	1939	354	486	0,0296	0,0054	0,0074
	Встречаемость в тексте в %				2,96%	0,54%	0,74%

Табл. 15. Анализ встречаемости в тексте произведения «Эхо орбитального удара» М. Глебова частиц «не», «бы», «же» Статистика: количество слов 65500 (текст сокращен)

№	Словарный объем текста N (число слов в документе, статистика)	Частота появления частиц			Весовые коэффициенты частиц (K)		
		не	бы	же	$K_{не}$	$K_{бы}$	$K_{же}$
1	65500	1674	205	132	0,0255	0,0031	0,0020
	Встречаемость в тексте в %				2,55%	0,31%	0,20%

3-й текст М. Глебова «Безусловная директива»

Табл. 16. Анализ встречаемости в тексте произведения «Безусловная директива» М. Глебова предлогов «в», «на», «с». Статистика: количество слов 65500 (текст сокращен)

№	Словарный объем текста N (число слов в документе, статистика)	Частота появления предлогов			Весовые коэффициенты предлогов (K)		
		в	на	с	$K_в$	$K_{на}$	$K_с$
1	65500	1596	1091	730	0,0243	0,0166	0,0111
	Встречаемость в тексте в %				2,43%	1,66%	1,11%

Табл. 17. Анализ встречаемости в тексте произведения «Безусловная директива» М. Глебова союзов «и», «а», «но». Статистика: количество слов 65500 (текст сокращен)

№	Словарный объем текста N (число слов в документе, статистика)	Частота появления союзов			Весовые коэффициенты союзов (K)		
		и	а	но	$K_и$	$K_а$	$K_{но}$
1	65500	1880	438	474	0,0287	0,0066	0,0072
	Встречаемость в тексте в %				2,87%	0,66%	0,72%

Табл. 18. Анализ встречаемости в тексте произведения «Безусловная директива» М. Глебова частиц «не», «бы», «же». Статистика: количество слов 65500 (текст сокращен)

№	Словарный объем текста N (число слов в документе, статистика)	Частота появления частиц			Весовые коэффициенты частиц (K)		
		не	бы	же	$K_{не}$	$K_{бы}$	$K_{же}$
1	65500	1641	222	119	0,0250	0,0033	0,0018
	Встречаемость в тексте в %				2,50%	0,33%	0,18%

4. Результаты сопоставления текстов

При сопоставлении этих табличных данных с использованием полученного по произведениям М. Глебова показателя идентичности, было установлено следующее. Первое равенство однотипных коэффициентов первого произведения («Столица мятежной окраины», Табл.10) и третьего произведений («Безусловная директива», табл.16), соответственно, $k_{1в} = k_{3в} = 0$ % (0,0243 =0,0243; второе равенство однотипных коэффициентов первого произведения (Табл.12) и второго произведений («Эхо орбитального удара», Табл. 15), соответственно, $k_{1бы} = k_{2бы} = 0$% (0,0031=0,0031); разность – 0,02% однотипных весовых коэффициентов первого произведения (Табл. 12) и второго произведения (Табл. 15), соответственно, $k_{1же}$ и $k_{2же}$. Однотипные весовые коэффициенты второго и третьего произведений составляют, соответственно, $k_{2но}$ и $k_{3но}$, а весовые коэффициенты третьего произведения, второго и первого произведения, соответственно, $k_{3бы}$, $k_{2бы}$ и $k_{1бы}$. Повторяемостью называется совпадение весовых коэффициентов, сравнением весовых коэффициентов устанавливается показатель совпадений их числовых значений. Из приведенных выше расчетов можно заключить, что показатель идентичности в трех произведениях Глебова, изменяющегося от 0 до 0,02 процентов, составляет 5.

Разброс по показателям идентичности в произведениях С. Лукьяненко и М. Глебова может быть объяснен тем, что С. Лукьяненко в своих произведениях (первый текст, Табл.1) чаще всего использует союз «и» (3,05 %). Тогда как по произведениям М. Глебова процентный показатель частоты употребления этого же служебного слова составляет 2,96 % (второе произведение, Табл.14). По использованию предлога «с» у обоих авторов наблюдается почти равенство (Лукьяненко – 1,03 %, Глебов – 1,04%. По использованию других служебных слов: например, наблюдается равенство частиц «бы» второго

произведения Лукьяненко (0,31%) с первым и вторым произведениями Глебова (0,31%). Максимальная разница (0,36 %) была выявлена по использованию союза «и», который у Лукьяненко – 3,05 % (первое произведение, Табл.2), а у Глебова – 2,79 % (первое произведение, Табл.11).

Этот сравнительный анализ по использованию служебных слов в произведениях указанных авторов может быть продолжен. Однако приведенные выше примеры, по мнению авторов данной работы, будут достаточны, чтобы констатировать особый идиостилевой авторский профиль и выявить некоторые особенности текста. То есть, указанные нами маркеры в целом характеризуют стилевые черты каждого автора при составлении их текстов.

Как показали проведенные исследования шести произведений двух авторов, показатель идентичности (выявленный на основе совпадения, повторяемости числовых значений) по трем произведениям С. Лукьяненко равен 3 с диапазоном изменения от 0 до 0,03%, а показатель идентичности по трем произведениям М. Глебова равен 5 с диапазоном изменения от 0 до 0,02%.

Далее, путем сопоставления однотипных весовых коэффициентов по заведомо известным текстам произведений находится среднее значение показателей идентичности шести указанных произведений. Не трудно догадаться, что в этом случае, среднее значение этих показателей составляет 4 с диапазоном изменения от 0 до 0, 03%.

Итак, критерий оценки, не отвечающий проверке на авторскую уникальность текстов указанных произведений, должен быть равным 4, и может изменяться от 0 до 0,03 %. По приведенным в таблицах вычислительным результатам видно, что первое равенство однотипных коэффициентов второго произведения С. Лукьяненко и первого произведения М. Глебова $k_{2бы}=k_{1бы}=0\%$. Второе равенство второго произведения С. Лукьяненко и второго произведения М. Глебова $k_{2бы} = k_{2бы} = 0\%$. Третье равенство третьего произведения С. Лукьяненко и второго произведения М. Глебова $k_{3на} = k_{2на} = 0\%$.

Далее сравнительный анализ показал, что первая разность 0,01% однотипных коэффициентов первого произведения С. Лукьяненко и первого произведения М. Глебова $k_{1с}$ и $k_{1с}$, а вторая разность 0,01% третьего произведения С. Лукьяненко и второго произведения М. Глебова $k_{3же}$ и $k_{2же}$. Тогда как третья разность равна 0,01% третьего текста С. Лукьяненко и второго текста произведения М. Глебова, соответственно, $k_{3и}$ и $k_{2и}$. Далее первая разность равна 0,02% второго произведения С. Лукьяненко и первого произведения М. Глебова $k_{2а}$ и $k_{1а}$. Вторая разность второго произведения С. Лукьяненко и третьего произведения М. Глебова $k_{2бы}$ и $k_{3бы}$ равна 0,02%. Первая разность первого произведения Лукьяненко и первого произведения Глебова $k_{1же}$ и $k_{1же}$ равна 0,03%, тогда как вторая разность равна 0,03% для первого произведения Лукьяненко и второго произведения Глебова $k_{1на}$ и $k_{2на}$. Третья разность первого произведения Лукьяненко и третьего произведения Глебова $k_{1а}$ и $k_{3а}$ равна 0,03%. Тогда как четвертая разность равна 0,03% для второго произведения Лукьяненко и первого произведения Глебова $k_{2с}$ и $k_{1с}$. Пятая разность равна 0,03% второго произведения Лукьяненко и первого произведения Глебова, соответственно, $k_{2же}$ и $k_{1же}$. Шестая разность 0,03% второго произведения Лукьяненко и второго произведения Глебова $k_{2и}$ и $k_{2и}$; а седьмая разность 0,03% третьего произведения Лукьяненко и второго произведения Глебова $k_{3но}$ и $k_{2но}$, и восьмая разность 0,03% третьего произведения Лукьяненко и третьего произведения Глебова соответственно $k_{3же}$ и $k_{3же}$.

Из представленных вычисленных результатов следует, что показатель идентичности по параметру атрибутивности текстов по одному условию (диапазона его изменения от 0 до 0, 03%) удовлетворяет требованиям на авторскую уникальность текста, но по второму условию критерия (не более 4), противоречит указанным требованиям, так как его повторяемость во всех шести произведениях составляет 16. Следовательно, рассматриваемые выше произведения не могут принадлежать одному автору.

Иными словами, согласно полученным вычислительным результатам, для подтверждения авторства (подтверждения идиостилевого профиля автора, формализованная проверка текста на авторскую уникальность) необходимым является выполнение одновременно двух условий изменения показателя идентичности вопреки корреляции между сравниваемыми произведениями двух или более авторов. Подтверждение

того, что результаты проведенных исследований могут быть распространены на установление авторства текстов нескольких произведений неизвестных (и даже искусственно созданных ChatGPT) и известных авторов с условием наличия хоть одного произведения известного автора как эталонного (образцовый текст) требует дальнейших исследований. В планы авторов входит разработка полного комплекса частных вычислительных методов лингвистического анализа текстов.

Заключение

В исследовании предложена методика оценивания литературно-художественных текстов по наиболее частотным (определенных как маркеры) служебным частям речи, включающая определение их весовых коэффициентов, вычисляемых отношениями количества соответствующих служебных слов в тексте к общему объему рассматриваемого текста.

Приведены и проанализированы результаты практических исследований, связанные с вычислением показателей идентичности у равных по объему и жанру текстов по однотипным весовым коэффициентам для двух авторов с целью выявления лексических особенностей произведений и уточнения идеостилевого авторского профиля.

При сопоставлении текстов (с сокращенным объемом слов для корректности сравнения) художественных произведений двух авторов выявлены характерные признаки в виде лексических показателей идентичности с диапазонами колебания их весовых коэффициентов, определяющих степень близости текстов разных авторов по стилю и некоторым лексическим особенностям.

Литература

1. Михеев М.Ю., Эрлих Л.И. Идиостилевой профиль и определение авторства текста по частотам служебных слов // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2018. №2. С. 25-34
2. Михеев М.Ю., Эрлих Л.И. Текстовые скрепы и их частоты как различительный признак авторских идиостилей Электронный ресурс. Дата обращения 15.03.2024. https://ruslang.ru/sites/default/files/doc/grigoriev2022/Mish.pdf
3. Орлова, М. В**.** Михеев М.Ю., Эрлих Л.И. Об отличиях русского научного идиостиля от художественного по частотам употребления текстовых скреп / Вопросы литературы. 2022. № 1. с.118-140.
4. Кукушкина О.В., Поликарпов А.А. Частотные и распределительные характеристики русских предлогов и синтаксем, с ними связанных (по «Ядерному компьютерному корпусу текстов русских газет конца XX-ого века») // Язык, сознание, коммуникация. Выпуск 47. М.: МАКС Пресс. том 47. 2013. с. 341-362
5. Всеволодова М.В., Кукушкина О.В., Поликарпов А.А. Русские предлоги и средства предложного типа. Материалы к функционально-грамматическому описанию реального употребления. Книга 1. Введение в объективную грамматику и лексикографию русских предложных единиц. М.: URSS, 2013. 304 с.
6. Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ данных. М.: НИУ ВШЭ. 2017. 269 с.
7. Смирнов И.В., Шелманов А.О. Семантико-синтаксический анализ естественных языков // Искусственный интеллект и принятие решений. 2013. № 1. С. 41-54.
8. Морозов Н.А. Лингвистические спектры: Средство для отличения плагиатов от истинных произведений того или другого известного автора: Стилеметрический этюд // Известия Отдела русского языка и словесности Императорской Академиинаук. 1915. Т. 20, кн. 4. С. 93-127.
9. Марков А.А. Об одном применении статистического метода // Изв.Имп.акад.наук. Серия 6. 1916. N4. с. 239-242.
10. Атрибуция текста: теория и практика (festivalnauki.ru). Электронный ресурс. Дата обращения 12.02.2024.
11. Smith Peter, Aldrigde W. (2011): Improving Authorship Attribution: Optimizing Burrows's Delta Method // Journal of Quantitative Linguistics 18. Pp. 63–88.
12. Рыжкович А.Ч. К вопросу о треминологическом статусе предлога // Universum: Филология и искусствоведение: электронный научный журнал. 2018. № 9(55). URL: http://7universum.com/ru/ philology/archive/item/6385
13. Shalymov D, Granichin O, Klebanov L, Volkovich Z. Literary writing style recognition via a minimal spanning tree-based approach // Expert Systems with Applications 61, 145-153, 2016 DOI: 10.1016/j.eswa.2016.05.032
14. Суетин В. Ю. Применение частотных характеристик для определения авторства литературных текстов // Вестник ТвГУ. Серия: Прикладная математика. 2022, выпуск 2. с. 84–89 DOI: 10.26456/vtpmk637
15. Воронина, М. Ю. Орлов Ю. Н. Определение автора текста методом сегментации. Федеральный исследовательский центр «Институт прикладной математики им. М. В. Келдыша Российской академии наук», 2022. DOI: https://doi.org/10.20537/2076-7633-2022-14-5-1199-1210
16. Кислицын А.А., Кислицына М.Ю. Распознавание выборочных распределений среди системы эталонов: метод ближайшего соседа // Препринты ИПМ им. М.В.Келдыша. 2023. № 29. 21с. https://doi.org/10.20948/prepr-2023-29 https://library.keldysh.ru/preprint.asp?id=2023-2
17. Горожанов А. И. Создание лингвистического корпуса на основе инструментов обработки естественного языка: планирование программных решений // Филологические науки. Вопросы теории и практики. 2023. Том 16. Выпуск 5. С. 1616-1620.

18. Гаспаров М.Л. Лингвистика стиха // Известия РАН. Сер. лит. и языка. – М., 1994. Т. 53. № 6. с. 28-35.
19. Федянин Д. Н., Русяева Е. Ю., Ахобадзе Г. Н. Лингвистический анализатор текста: Свидетельство о государственной регистрации программы для ЭВМ № 2023668307 РФ; Зарег. 25.08.2023.
20. Литрес. Электронный ресурс: https://www.litres.ru/book/maks-alekseevich-glebov/chernyy-staratel-67077536/ Дата обращения 10.01.2024.

Русяева Елена Юрьевна. Федеральное государственное бюджетное учреждение науки Институт проблем управления им. В. А. Трапезникова Российской академии наук. Старший научный сотрудник, кандидат философских наук. Область научных интересов: философия управления, информационные технологии лингвистического анализа текстов, когнитивное моделирование, концептуализация вычислительных моделей ИАС. E-mail: rusyaeva@ipu.ru

Ахобадзе Гурами Николаевич. Федеральное государственное бюджетное учреждение науки Институт проблем управления им. В. А. Трапезникова Российской академии наук. Ведущий научный сотрудник, доктор технических наук. Область научных интересов: вычислительная техника, информационные технологии ИАС. E-mail: ahogur@ipu.ru

Some Features of Literary Texts when Comparing them to Determine their Authorship

G. N. Akhobadze, E. Yu. Rusyaeva

V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russia

Abstract. A method for analyzing literary author's texts based on selecting the most frequent auxiliary parts of speech characteristic of a particular author's style and calculating their weighting coefficients has been developed. This linguistic analysis of natural language text (NLP) is based on the calculation of the most frequently used prepositions, conjunctions and particles in literary works. The process of calculating weight coefficients, determined by the ratio of the values of auxiliary parts of speech in the text to its total volume, has been analyzed in detail. Experimental results on establishing the authorship of literary texts for two authors are presented. The results were obtained by comparing the numerical values of the same type of weighting coefficients, expressed as percentages. The theoretical and practical results obtained can be used to analyze, identify linguistic features, and differences not only in literary texts, but, in the future, in texts of any genre and style.

Keywords: weight coefficient, auxiliary part of speech, authorship, text, identity indicator, repeatability.

DOI 10.14357/20718632240207 **EDN** WGWOTS

References

1. Mikheev M.Yu., Erlikh L.I. Idiostyle profile and determination of the authorship of the text by the frequencies of function words // Scientific and technical information. Series 2: Information processes and systems. No. 2, 2018. – p. 25-34
2. Mikheev M.Yu., Erlikh L.I. Text staples and their frequencies as a distinctive feature of the author's idiostyles Electronic resource. Access date 03/15/2024. https://ruslang.ru/sites/default/files/doc/grigoriev2022/Mish.pdf
3. Orlova, M.V. Mikheev M.Yu., Erlikh L.I. On the differences between Russian scientific idiostyle and artistic style in terms of the frequency of use of text staples / Questions of literature. No. 1 2022, pp. 118-140.
4. Kukushkina O.V., Polikarpov A.A. Frequency and distribution characteristics of Russian prepositions and syntaxes associated with them (according to the "Nuclear Computer Corpus of Texts of Russian Newspapers of the End of the 20th Century") // Language, consciousness, communication. Issue 47. M.: MAKS Press, volume 47, 2013. .341-362
5. Vsevolodova M.V., Kukushkina O.V., Polikarpov A.A. Russian prepositions and prepositional devices. Materials for a functional-grammatical description of real use. Book 1. Introduction to objective grammar and lexicography of Russian prepositional units. Place of publication URSS, 2013. 304 p.

6. Bolshakova E.I., Vorontsov K.V., Efremova N.E., Klyshinsky E.S., Lukashevich N.V., Sapin A.S. Automatic processing of texts in natural language and data analysis - M.: National Research University Higher School of Economics, 2017. - 269 p.
7. Smirnov I.V., Shelmanov A.O. Semantic-syntactic analysis of natural languages // Artificial intelligence and decision making. - 2013. No. 1. - P. 41-54.
8. Morozov N.A. Linguistic spectra: A means for distinguishing plagiarism from the true works of one or another famous author: Stylemetric study // News of the Department of Russian Language and Literature of the Imperial Academy of Sciences. - 1915. - T. 20, book. 4. - pp. 93-127.
9. Markov A.A. On one application of the statistical method//Izv.Imp.academy.nauk, Ser. 6. 1916. N4, pp. 239-242.
10. Text attribution: theory and practice (festivalnauki.ru). Electronic resource. Access date 02/12/2024.
11. Smith Peter, Aldrigde W. (2011): Improving Authorship Attribution: Optimizing Burrows's Delta Method // Journal of Quantitative Linguistics 18, 63–88.
12. Ryzhkovich A.Ch. On the question of the treminological status of the preposition // Universum: Philology and art history: electronic scientific journal. 2018. No. 9(55). URL: http://7universum.com/ru/philology/archive/item/6385
13. Shalymov D, Granichin O, Klebanov L, Volkovich Z. Literary writing style recognition via a minimal spanning tree-based approach // Expert Systems with Applications 61, 145-153, 2016 DOI: 10.1016/j.eswa.2016.05.032
14. Suetin V. Yu., Application of frequency characteristics to determine the authorship of literary texts, Vestnik TvGU. Series: Applied Mathematics, 2022, issue 2, 84–89 DOI: 10.26456/vtpmk637
15. Voronina, M. Yu. Orlov Yu. N. Determining the author of a text using the segmentation method. Federal Research Center "Institute of Applied Mathematics named after. M. V. Keldysh Russian Academy of Sciences", 2022. DOI: https://doi.org/10.20537/2076-7633-2022-14-5-1199-1210
16. Kislitsyn A.A., Kislitsyna M.Yu. Recognition of sample distributions among a system of standards: the nearest neighbor method // Preprints of IPM im. M.V.Keldysh. 2023. No. 29. – 21 p. https://doi.org/10.20948/prepr-2023-29 https://library.keldysh.ru/preprint.asp?id=2023-2
17. Gorozhanov A. I. Creation of a linguistic corpus based on natural language processing tools: planning software solutions. / Philological sciences. Questions of theory and practice., 2023. Volume 16. Issue 5. P. 1616-1620.
18. Gasparov M.L. Linguistics of verse // News of PAH. Ser. lit. and language. – M., 1994. – T. 53. No. 6. – p. 28-35.
19. Fedyanin D. N., Rusyaeva E. Yu., Akhobadze G. N. Linguistic text analyzer: Certificate of state registration of a computer program No. 2023668307 RF; Registered 08/25/2023.
20. Litres. Electronic resource: https://www.litres.ru/book/maks-alekseevich-glebov/chernyy-staratel-67077536/ Date of access 01/10/2024

Rusyaeva Elena Ya. PhD, V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences 65 Profsoyuznaya street, Moscow 117997, Russia. Research interests: philosophy of management, information technologies of linguistic analysis of texts, cognitive modeling, conceptualization of computational models of IAS. E-mail: rusyaeva@ipu.ru

Akhobadze Gurami N. Professor, Doctor of technical sciences, V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences 65 Profsoyuznaya street, Moscow 117997, Russia. Area of scientific interests: computer engineering, information technologies IAS. E-mail: ahogur@ipu.ru