

Стилистические сходства и различия русских межвоенных романов через призму наиболее частотной лексики

Федор Никитич Двинятин¹, Борис Вадимович Ковалев², Санкт-Петербургский
государственный университет (Россия, Санкт-Петербург)^{1,2}, f.dvinyatin@spbu.ru¹; bvkovalev@yandex.ru²

DOI: 10.31857/S0131611725060082

АННОТАЦИЯ: Статья посвящена анализу наиболее частотной лексики в корпусе русской межвоенной прозы. Отправной точкой является стилиметрический анализ при помощи одного из самых апробированных и надежных на сегодняшний день стилиметрических инструментов — Delta Берроуза. С его помощью была осуществлена кластеризация романов межвоенного периода. В рамках настоящей статьи предлагается провести анализ наиболее частотной лексики в этих романах и сопоставить результаты с данными Delta. Особое внимание уделяется грамматическим категориям, изучению признаков, непосредственно выраженных в частотном словоупотреблении. Цель: выявить, что частотность этих признаков может сказать о стилистических особенностях анализируемых текстов, их сходствах и различиях. Для этого вводятся понятия высокого, нейтрального и низкого уровня частотности некоторого слова в определенном тексте. На основании подсчетов обнаруживается, что резко маркированы стили Шмелева и Пильняка, для всех 25 слов уровень частотности совпадает в «Дороге на океан» и «Скутаревском» Леонова. Результаты позволяют зафиксировать конкуренцию предлогов, союзов и местоимений. Ни в одном из рассмотренных текстов нет сочетания одинаково высокой или одинаково низкой доли предлогов *в* и *на*. Кроме того, высокая частотность как

соотносится с метафорической поэтикой Олеши, пяти более поздних романов Набокова, Федина, Шмелева и Грина; обилие местоимения *ты* маркирует тексты с большой ролью диалога и неформальным, неофициальным характером этого диалога (Гайдар, Леонов, Платонов). Частотность местоимения *она* глобально сопряжена со значимой ролью героинь в повествовании, но обнаруживается, что уровень частотности *она* для «Блестящего мира» существенно выше, чем для «Бегущей по волнам».

КЛЮЧЕВЫЕ СЛОВА: количественная грамматика, стилеметрия, Delta Берроуза, наиболее частотная лексика, русская межвоенная проза

для цитирования: Двинятин Ф. Н., Ковалев Б. В. Стилистические сходства и различия русских межвоенных романов через призму наиболее частотной лексики // Русская речь. 2025. № 6. С. 111–127. DOI: 10.31857/S0131611725060082.

The Language of Fiction

Stylistic Similarities and Differences of Russian Interwar Novels Through the Prism of the Most Frequent Words

Fyodor N. Dviniatin¹, Boris V. Kovalev², Saint Petersburg State University
(Russia, Saint Petersburg)^{1,2}, f.dvinyatin@spbu.ru¹; bvkovalev@yandex.ru²

ABSTRACT: The article deals with the analysis of the most frequent words in the corpus of Russian interwar prose. The starting point is a stylistic analysis using one of the most tested and reliable stylistic tools, Burrows's Delta, with the help of which we performed the clustering of the interwar period novels. Within the framework of this article, it is proposed to analyze the most frequent words in these novels and compare the results with the Delta

data. Particular attention is paid to grammatical categories, the study of features directly expressed in frequency of word usage. The objective is to identify what the frequency of these features can say about the stylistic features of the analyzed texts; their similarities and differences. For this purpose, we introduced the concepts of high, neutral and low frequency of a word in a certain text. Based on the calculations, it is found that the styles of Shmelev and Pilnyak are sharply marked, for all 25 words the frequency level coincides in Leonov's "Road to the Ocean" and "Skutarevsky". The results also allow us to record the competition of prepositions, conjunctions and pronouns. None of the texts examined has a combination of equally high or equally low proportions of the prepositions "в" and "на". In addition, it is concluded that the high frequency of "как" correlates with the metaphorical poetics of Olesha, five later novels by Nabokov, Fedin, Shmelev and Green, and that the abundance of the pronoun "ты" marks texts with a large role of dialogue and the informal, unofficial nature of this dialogue (Gaidar, Leonov, Platonov). The frequency of the pronoun "она" is globally associated with the high role of the heroines in the narrative, but it turns out that the level of frequency for "The Shining World" is significantly higher than for "Running on the Waves".

KEYWORDS: quantitative grammar, stylometry, Burrows's Delta, the most frequent words, Russian interwar prose

FOR CITATION: Dviniatin F. N., Kovalev B. V. Stylistic Similarities and Differences of Russian Interwar Novels Through the Prism of the Most Frequent Words. Russian Speech = Russkaya Rech'. 2025. No. 6. Pp. 111–127. DOI: 10.31857/S0131611725060082.

Введение

Изучение наиболее частотной лексики в контексте стилеметрической идентификации своеобразия авторского стиля того или иного писателя — одно из наиболее актуальных и быстро развивающихся направлений на сегодняшний день.

Первые примеры анализа наиболее частотной лексики наблюдаются еще на начальном этапе использования квантитативных методов в филологии [Lutoslawski 1897; Морозов 1915]. Это направление активно развивалось в середине XX в. как в англоязычной [Mosteller, Wallace 1963], так и во франкоязычной лингвистической традиции [Michéa 1953].

Сегодня работы подобного типа встречаются регулярно, различаются методики и материал. Так, в Воронеже существует школа маркемологического анализа, которая предполагает изучение лексики с точки зрения частотности и длины слов [Кретов 2009; Кретов, Фаустов 2017]. Методика доказала эффективность на материале русскоязычной поэзии и прозы, однако применимость метода к англоязычным текстам [Артемова 2017] пока не представляется очевидной [Кашкина 2010; Зевахина, Филиппова 2024]. Квантитативный анализ разных уровней текста (от цветообозначений до морфологических и ритмо-метрических параметров) используется и в рамках Смоленской школы [Андреев 2012; Андреев, Павлова, Романова 2021]; частотные текстовые «скрепы» на материале служебной лексики для определения идиостилевого профиля автора выявляются, например, в: [Михеев, Эрлих 2018].

Однако нельзя не признать, что наиболее популярным и апробированным инструментом, основанным на подсчете наиболее частотной лексики, является Delta Берроуза [Burrows 2002]. Для каждого из анализируемых текстов нужно взять некоторое количество самых частотных слов, а затем сравнить их частотности (представленные как мера z-score, учитывающая число всех слов в тексте), суммируя разницу (т. е. дельту) для каждой пары текстов. Основная закономерность формулируется следующим образом: чем это расстояние меньше, тем более стилистически сходными оказываются тексты. В фокусе анализа оказывается по большей части служебная лексика.

Главными ограничителями для анализа выступают требования жанровой гомогенности (например, стихи, художественная проза и драматургия одного и того же автора характеризуются разными авторскими сигналами) и достаточного объема текста (от 5000 слов).

Алгоритм доказал свою эффективность на материале разных языков: от испанского [Calvo Tello 2019] до польского [Eder 2011], от латыни [Rybicki, Eder 2011] до русского [Орехов 2021; Двинятин, Ковалев 2024].

Пожалуй, наиболее ярким примером применения Delta можно назвать следующий. Допустим, исследователь хочет использовать этот инструмент для формулирования атрибуционной гипотезы, то есть формализованного предположения о том, кто в действительности является автором некоторого текста. Тогда необходимо поместить в корпус анализируемый текст (A_x), приписываемый некоторому автору A, а также еще некоторое количество соразмерных (обязательно более 5000 слов) текстов ($A_1, A_2...$), написанных тем же автором примерно в тот же хронологический промежуток. Также нужно добавить в корпус произведения иных авторов-претендентов (B_1, B_2, C_1, C_2 и т. д., подбор делается на

базе литературоведческих наблюдений и высказываемых умозрительных гипотез, основанных на историко-литературных и биографических данных). Таких текстов должно быть как минимум по два на каждого автора. Если атрибутируемый текст (A_x) кластеризуется (то есть помещается алгоритмом на основании кратчайшего межтекстового расстояния) на одной ветке с другими текстами того же предполагаемого автора A , значит, эти тексты в рамках настоящего корпуса оцениваются алгоритмом как стилистически близкие, следовательно, можно предположить, что они написаны одним автором, что далее необходимо подтверждать на архивном, историко-литературном уровне.

Итак, Delta активно используется в современной цифровой филологии как вспомогательный инструмент для процедуры атрибуции, для решения задач периодизации. Однако, на наш взгляд, одно из наиболее перспективных направлений — сравнительно-сопоставительное изучение стилистических особенностей текстов, написанных разными авторами в конкретный период времени. Это позволяет ставить вопрос о стилеметрическом срезе литературного языка, используемого авторами определенной эпохи в определенном жанре.

Особый интерес для нас представляет анализ русской межвоенной прозы, творчество крупнейших романистов 1920–1940-х гг. Мы исследовали [Двинятин, Ковалев 2024] этот пласт текстов при помощи Delta, подвергнув результаты историко-филологической интерпретации; наше нынешнее исследование фокусируется на лингвистических признаках, выражающихся в наиболее частотной лексике, которые способствуют либо объединению, либо различению текстов.

Цель состоит в том, чтобы в первом приближении выявить, что частотность употребления этих малых признаков может сказать нам о стилистических особенностях анализируемых текстов, их сходствах и различиях. Представляется важным обратиться не только к сугубо лексическому уровню, но и к грамматике; не только к собственно частотному словоупотреблению, но и к тому, что в нем непосредственно отражено. Иными словами, мы намерены опираться на грамматические категории, выраженные в наиболее частотной лексике, для попытки объяснительного анализа, поскольку служебная лексика может считаться скрытым признаком распределений грамматических категорий и конструкций.

Постановка эксперимента

В исследовательский корпус были включены следующие тексты (Табл. 1). Принципиально, что лемматизация не проводилась. Эксперименты проводились в R, пакет Stylo [Eder, Rybicki et al. 2016].

Таблица 1. Состав и объем исследуемых текстов
Table 1. Analyzed texts and their volume

Автор	Текст	Год создания текста ¹	Объем словоупотреблений
В. Набоков	«Машенька»	1926	25 759
В. Набоков	«Защита Лужина»	1929–1930	53 015
В. Набоков	«Король, дама, валет»	1928	56 889
В. Набоков	«Камера обскура»	1931	44 851
В. Набоков	«Подвиг»	1932	50 316
В. Набоков	«Приглашение на казнь»	1935	39 759
В. Набоков	«Отчаяние»	1934	43 874
В. Набоков	«Дар»	1938	105 454
К. Вагинов	«Труды и дни Свистонова»	1929	26 937
К. Вагинов	«Козлиная песнь»	1928	40 464
Г. Газданов	«Вечер у Клэр»	1929	44 739
Г. Газданов	«Ночные дороги»	1941	64 798
Г. Газданов	«Призрак Александра Вольфа»	1939–1946	38 230
Г. Газданов	«Полет»	1939	60 704
А. Гайдар	«Военная тайна»	1935	30 229
А. Гайдар	«Судьба барабанщика»	1939	28 180
А. Грин	«Бегущая по волнам»	1928	56 480
А. Грин	«Блистающий мир»	1923	47 536
Б. Зайцев	«Дом в Пасси»	1935	45 818
Б. Зайцев	«Золотой узор»	1926	63 420
М. Зощенко	«Голубая книга»	1935	84 447
М. Зощенко	«Сентиментальные повести»	1923–1930	50 263
И. Ильф и Е. Петров	«Двенадцать стульев»	1927	102 554
И. Ильф и Е. Петров	«Золотой теленок»	1931	91 768
В. Катаев	«Белеет парус одинокий»	1936	70 216
В. Катаев	«Время, вперед»	1932	72 402

¹ Для «составных» текстов и произведений со сложной текстологией указаны интервалы, в которые были написаны соответствующие тексты, в большинстве случаев приведены даты публикации полной версии.

Окончание табл. 1

Автор	Текст	Год создания текста ¹	Объем словоупотреблений
Л. Леонов	«Барсуки»	1924	102 297
Л. Леонов	«Скутаревский»	1932	95 462
Л. Леонов	«Соть»	1930	90 732
Л. Леонов	«Дорога на океан»	1935	161 788
Ю. Олеша	«Зависть»	1927	29 505
Ю. Олеша	«Три толстяка»	1924	29 381
Б. Пильняк	«Голый год»	1922	42 210
Б. Пильняк	«Красное дерево»	1929	12 563
А. Платонов	«Котлован»	1930	35 771
А. Платонов	«Ювенильное море»	1934	23 436
А. Н. Толстой	«Петр Первый»	1929–1944	228 808
А. Н. Толстой	«Хожение по мукам» (1, 2, 3)	1920–1941	94 595 + 89 702 + + 109 030
Ю. Тынянов	«Кюхля»	1925	84 144
Ю. Тынянов	«Смерть Вазир Мухтара»	1928	115 774
К. Федин	«Братья»	1926–1928	107 267
К. Федин	«Города и годы»	1921–1924	100 109
И. Шмелев	«Богомолье»	1931	42 068
И. Шмелев	«Лето Господне»	1934–1944	121 350
Итого			3 155 094

Результаты кластеризации представлены на дендрограмме (Рис. 1).

Читается график следующим образом. Наиболее стилистически близкие тексты соединяются как своего рода «листья» на ветке; другие, несходные с предыдущими, но похожие между собой, объединяются на другой ветке. Чем меньше межтекстовое расстояние Delta (схематично обозначено на шкале внизу рисунка), тем ближе располагаются тексты, что дает основания говорить о них как о стилистически сходных. Как видим, Delta безошибочно относит тексты одного автора в одну ветвь (единственное исключение — Олеша и его «Три толстяка» и «Зависть» — объясняется разной жанровой природой текстов, а к этому показателю алгоритм чувствителен). Подробное обсуждение результатов замеров и их интерпретация в историко-филологическом ключе приведены в статье [Двинятин, Ковалев 2024].

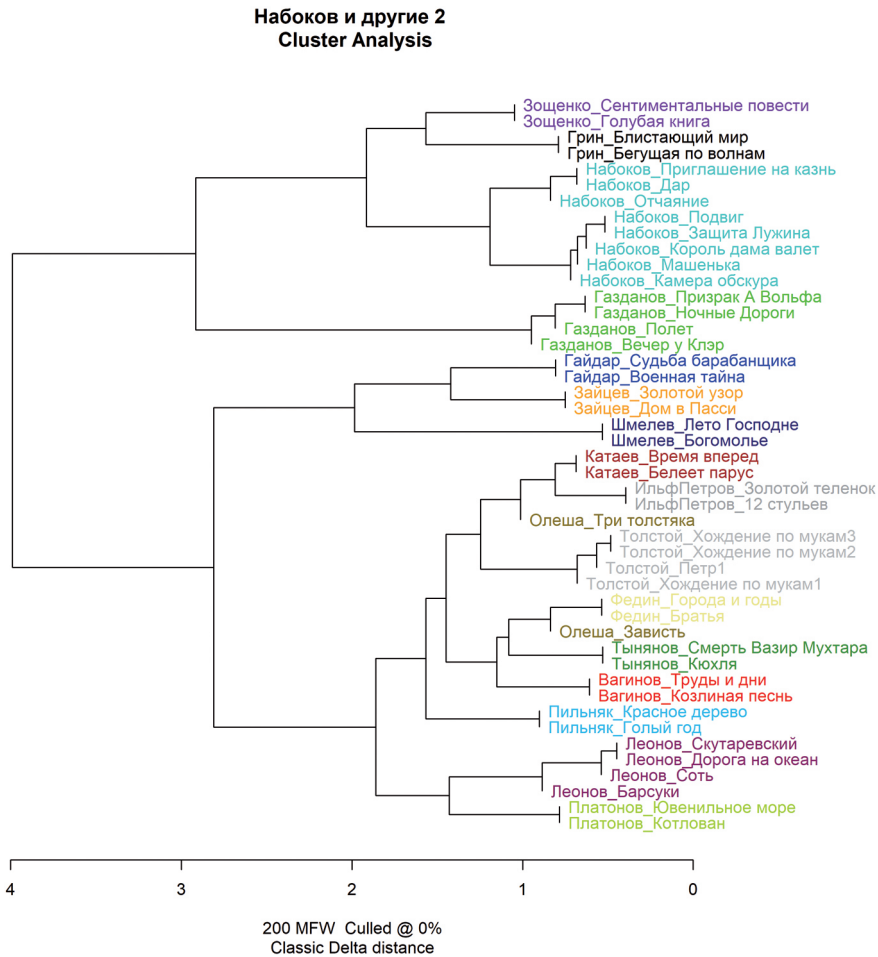


Рис. 1. Кластеризация романов межвоенной поры посредством метода Delta. 200 наиболее частотных слов

Fig. 1. Clustering of Interwar Novels Using Delta Method. 200 Most Frequent Words

Далее мы сосредоточимся на анализе 25 наиболее частотных слов (подчеркнем, что речь идет об относительных частотах) по романам, помещенным в корпус. Наша задача — проанализировать специфику наиболее частотной лексики и на этой основе определить, за счет каких лингвистических признаков, выраженных в частотном словоупотреблении, тексты оказываются связаны между собой или, напротив, отличаются друг от друга, и сопоставить наблюдения с данными Delta.

Результаты и обсуждение

Мы вводим понятия высокого, нейтрального и низкого уровня частотности некоторого слова в определенном тексте. Эти уровни совершенно условны и имеют целью только предварительное сравнение текстов между собой, одновременно объективное и достаточно компактное.

Delta — инструмент формализованный и не объясняющий механизм своей работы (кроме внешних условий). Чтобы избежать процедурной тавтологии, нам казалось уместным при первых попытках объяснения отступить от жесткой формализации. Разброс по разным словоформам был различен, группы текстов с повышенной и пониженной частотностью для разных словоформ тоже обрисовывались различным образом. Для определения порогов высокой и низкой частотности был выбран баланс двух критериев: мера отступления от средней и медианной частотности данной словоформы во всех текстах корпуса и количество текстов, распределяемых такими порогами в верхнюю или нижнюю часть списка (по частотности данной словоформы). Двумя дополнительными критериями были плотность показателей в ряду частотности одного слова по разным текстам (правильным считалось разбивать порогом в месте определенного разрыва по частотностям, а не там, где разница между словами соседних рангов почти незаметна) и возможность приурочить порог к показателю, кратному одной десятой процента. Поначалу мы предполагали, что для каждой словоформы область высокой и область низкой частотности будут обнимать приблизительно по 20 % случаев, но материал подсказывал, что усредненная частотность, трудная для расчленения, начинается в среднем после отделения 15 % текстов с высокой либо низкой частотностью. Точно так же отклонение порогов высокой и низкой частотности от медианных значений изначально предполагалось около 25 %. В любом случае материал не подгонялся ни под какое конкретное решение: цель состояла не в доказательстве (или опровержении) какой-либо конкретной гипотезы, а в поиске первого контура объяснения.

Мы приняли такие пороги (знаки означают, что частотность выше такого уровня ↑ считалась высокой, ниже такого ↓ — низкой): *и*: 4,6↑, 3,4↓; *в*: 2,9↑, 2↓; *на* 1,9↑, 1,5↓; *не* 1,9↑, 1,5↓; *с* 1,4↑, 1↓; *он* 1,7↑, 0,8↓; *что* 1,7↑, 0,8↓; *я* 2↑, 0,5↓; *как* 0,9↑, 0,55↓; *а* 0,9↑, 0,4↓; *его* 0,8↑, 0,4↓; *то* 0,9↑, 0,4↓; *но* 0,65↑, 0,45↓; *к* 0,75↑, 0,45↓; *все* 0,7↑, 0,4↓; *у* 0,7↑, 0,4↓; *за* 0,6↑, 0,4↓; *но* 0,7↑, 0,4↓; *это* 0,8↑, 0,4↓; *из* 0,5↑, 0,325↓ (подавляющее большинство частотностей по этому предлогу располагается очень плотно друг к другу); *она* 0,7↑, 0,3↓; *было* 0,65↑, 0,325↓; *так* 0,5↑, 0,25↓; *от* 0,45↑, 0,25↓; *же* 0,45↑, 0,25↓.

В нашем корпусе обнаруживаются три автора, у которых в обоих представленных текстах уровень частотности совпадает для 22 из 25 слов

верхней части списка: это Шмелев, Вагинов, Ильф и Петров. Например, у Шмелева предлоги *на* (частотность для «Богомолья» 2,11451, для «Лета Господня» 2,15629; здесь и далее в процентах и с округлением до пятого знака после запятой), *с* (0,70726 и 0,67233), *у* (0,94221 и 0,84333), *от* (0,47716 и 0,47297), союзы *и* (5,16398 и 4,99992), *а* (2,15569 и 2,032), местоимение *все* (0,91072 и 1,00933), словоформы *то* (2,032 и 1,10442 — результат для «Богомолья» намного выше среднего по корпусу уровня, но результат по «Лету Господню» тоже основательно выше) и *как* (0,98338 и 0,97429) используются одинаково часто, то есть (намного) чаще среднего уровня; местоимения *он* (0,70726 и 0,67233), *что* (0,74359 и 0,67483), *она* (0,28097 и 0,29529), союз *но* (0,12111 и 0,16266), глагольная форма *было* (0,08235 и 0,10844), словоформа *же* (0,17924 и 0,16266) — одинаково редко; предлоги *в* (2,14116 и 2,31144), *по* (0,59827 и 0,5789), местоимения *я* (0,90103 и 0,91673), *его* (0,44325 и 0,37453), частица *не* (1,63978 и 1,61742) и словоформа *так* (0,51834 и 0,50133) — нейтрально по частоте; и только для предлогов *к* и *за* и местоимения *это* фиксируется расхождение: причем случай *к* более очевидный: редкий в «Лете Господнем» (0,44043), он нейтрален в «Богомолье» (0,591); в случаях *за* и *это* уровни в большей степени рассекла условно принятая граница: *за* 0,64186 в «Богомолье» (частый) и 0,53886 в «Лете Господнем» (нейтральный), *это* 0,41418 (нейтральный уровень) и 0,38955 (редкий) соответственно. Случай Шмелева показателен также и тем, что 15 из 22 совпадающих по частоте слов маркированы, т. е. вверх или вниз отличаются от средних показателей. Для Вагинова так же совпадают уровни частотности у 22 слов из 25 (частое: *по*, редкие: *не*, *было*, *так*, *же*, нейтральные: *и*, *на*, *с*, *что*, *я*, *как*, *а*, *его*, *то*, *к*, *все*, *у*, *за*, *но*, *из*, *она*, *от* — в отличие от Шмелева 17 из 22 совпадающих по уровню частоты — нейтральные); расходятся *в*, частое в «Козлиной песни» (2,9439) и нейтральное в «Трудах и днях Свистонова» (2,42333), *он* (1,71384 и 1,45246 соответственно — тоже частое и нейтральное), *это* (0,37313 и 0,46232 — наоборот, редкое и нейтральное). У Ильфа и Петрова еще больше совпадающих нейтральных — 19 из 22 (*в*, *на*, *с*, *он*, *что*, *я*, *как*, *а*, *его*, *по*, *к*, *все*, *у*, *за*, *но*, *это*, *из*, *она*, *было*, *от*, *же*), при одинаково повышенной в «Двенадцати стульях» и «Золотом теленке» частотности *из* и одинаково пониженной — *как* и *она*. Таковы особенно наглядные основания кластеризации текстов одного автора в одну ветвь.

В принципе в нашем материале есть два текста одного автора, для которых совпадает уровень частоты всех 25 верхних слов списка: это «Дорога на океан» и «Скутаревский» Леонова. При этом своеобразие этого случая в том, что уровень частоты для 24 из этих 25 слов — нейтральный (и повышенный — для *его*). Кроме того, есть еще два текста Леонова, и уровень

частотности для всех четырех текстов совпадает «только» в 20 из 25 случаев. При этом следует добавить, что при первой серии экспериментов в корпусе было только два романа Леонова: «Барсуки» и «Скутаревский» — и при таком составе алгоритм «смешивал» эти тексты с романами других авторов: «Скутаревского» с Набоковым, «Барсуков» — с А. Н. Толстым. И лишь после добавления «Дороги на океан» (чей частотный профиль схож со «Скутаревским») и «Соти» тексты Леонова выделились в отдельную самостоятельную ветвь.

В других случаях для верхней зоны списка частотностей (25 слов) число слов, совпадающих по условному уровню частотности, меньше. У Тынянова 21 такое слово (*в, на, не, с, он, что, я, а, его, то, по, к, все, у, за, но, это, она, было, от, же*), Катаева — 19, Федина и Грина — 18, Олеши — 17, Платонова, Зайцева, — 16, Гайдара, Зощенко — 15, Пильняка — 13, для Газданова — только 8, но для трех романов, кроме «Вечера у Клэр», 17, А. Н. Толстого — только 6 (*а, его, то, все, у, так*), но для трех частей «Хожdenия по мукам» — 16, в основном все нейтральные.

Кроме Шмелева, резко маркированы стили Пильняка (в одном тексте 14 слов, а в другом 16 попадают либо в редкие, либо в частые) и Толстого в «Петре Первом» (16 слов редких либо частых). Противоположный полюс, как сказано, Леонов. Другие случаи: ранний «Вечер у Клэр» Газданова (пять маркированных по частоте слов), 4 из 8 романов Набокова (от 2 до 4 маркированных слов, еще в двух по 6), обе вещи Олеши (6 и 4 маркированных слова). Заметим, что, во-первых, для сличения текстов между собой (например, одного автора) все равно, совпадают ли их частотности в верхней, нижней или нейтральной области, и, во-вторых, эти повышенные, пониженные или нейтральные частотности слов из верхней части списка не имеют непосредственного эстетического эффекта: Олеши и Набоков считаются превосходными стилистами, но и, наоборот, А. Н. Толстой и Шмелев тоже большие «мастера языка».

Ни один из текстов нашего корпуса не перекрывает частоту *на* в первой книге «Тихого Дона», известную по отдельному замеру на основании Национального корпуса русского языка (НКРЯ) — 2,3307 (максимально приближается «Петр Первый» — 2,26247). Точно так же «Москва под ударом» Андрея Белого (такой же отдельный замер) не превзойдена ни одним из текстов обсуждаемого корпуса по доле предлога *с* (как и в целом тексты ее автора) — 1,88536 (в данном эксперименте ближайшая доля у «Лета Господня» — 1,57405). Зато эксперимент позволил обнаружить автора, у которого предлога *в* оказалось еще больше, чем в «Белой гвардии» Булгакова (такой же отдельный замер), казавшейся до этого явным лидером в этом отношении: там 3,37967, а у Пильняка 3,66116 в «Голом годе»

и 3,51581 в «Красном дереве». В принципе, *в* — самый частый в русской речи предлог, начиная с древнейших текстов; превышение его частотности может говорить о реализации индивидуального стиля, сознательной или скорее бессознательной, не столько в поиске редких и оригинальных стиливых ресурсов, сколько в авторском усугублении общих тенденций; но это может говорить и об особом характере предпочитаемой микросемантики («одно в другом»), и просто о плохо семантизируемых языковых привычках. Относительная суммарная частота различных предлогов — очень надежное средство распределения стилей — слабо связана с общей семантической (в отличие от местоимений) и общей синтаксической (в отличие от союзов) перспективой текста: конкуренция предложной конструкции и адъективной конструкции, предложной конструкции и придаточного предложения, бесспорные примеры синтаксической роли предлогов, отражаются скорее в общей роли предлогов, чем в их соотношении, а связь предлогов с распределением синтаксических ролей в предложении и с построением синтаксических схем по образцу пространственных моделей, опять-таки бесспорная, пока, к сожалению, с трудом конкретизируется до учета относительной частотности отдельных предлогов.

Результаты позволяют зафиксировать ожидаемую, но все равно примечательную конкуренцию предлогов, союзов и местоимений. Ни в одном из рассмотренных текстов нет сочетания одинаково высокой или одинаково низкой доли предлогов *в* и *на* (проще и логичнее всего объяснить это именно конкуренцией: не возникает ситуации с «двумя победителями» или «двумя проигравшими»). При этом остальные 7 типов возможных сочетаний (нейтральный уровень для обоих предлогов и все 6 типов расхождения) фиксируются². Почти такое же положение с предлогами *с* и *из*, и все-таки не вполне: в трех текстах А. Н. Толстого из четырех их одинаково много. Так же взаимодополнительны союзы *а* и *но*: не бывают оба частыми или оба редкими; напротив, *а* и *и* могут совпадать в направлении отклонения: у Грина их одинаково мало, в «Военной тайне» Гайдара и у Шмелева одинаково много.

Почти во всех текстах не фиксируется одновременное повышение или понижение доли местоимений *он* и *она*, но и тут есть исключения: их одинаково много в трех романах Набокова (и если «Камера обскура» и «Король, дама, валет» — явные истории о мужчинах и женщинах, то для «Защиты Лукина» это более интригующе) и одинаково мало в «Голом годе»

² О конкуренции этих предлогов в древнерусском языке и о любопытнейшем случае их соотношения в «Слове о полку Игореве» см. со ссылкой на неопубликованный доклад одного из авторов настоящей статьи: [Введенский 2013].

Пильняка. Ожидаемую корреляцию между словоформами *он* и *его* проследить труднее, потому что *его* оказывается в нашем корпусе очень нейтральной по частотности словоформой: доли варьируют от ~0,3 до ~0,8, да и такие значения единичны, в основном около 0,5–0,7, тогда как *он* демонстрирует размах от ~0,2 до ~2,2 (ср. высокий порог — 1,7 — принятый нами для признания верхнего уровня частотности *он*, простейшее объяснение в том, что *его* часто используется в функции притяжательного местоимения, тематически более нейтрального, чем личное местоимение *он*, сильнее связанное с указанием на действующее лицо).

В нескольких случаях в стилистический замер отчетливо вторгаются смысловые параметры.

а) Доля местоимения *я* и перволичная повествовательная форма. Корреляция очень сильная: высокая доля *я* в «Судьбе барабанщика», но не в «Военной тайне»; в «Бегущей по волнам», но не «Блистающем мире»; в «Зависти», но не в «Трех толстяках»; в «Отчаянии», но ни в каком другом романе Набокова — все перволичные вещи (в перволичном «Лете Господнем» доля *я* тоже повышена, но не столь значительна, и это соотносится с тем, что *я*-рассказчик более перифериен).

б) Местоимение *она* и роль героинь. Хотя, вообще говоря, местоимение женского рода необязательно должно быть связано с персонажами женского пола⁵, все же причина резкого повышения его роли в текстах почти наверняка должна быть связана с ролью в действии героинь. У *она* высокий показатель в романах «Защита Лужина», «Камера обскура», «Король, дама, валет», низкий в «Петре Первом», но не во всех трех частях «Хождения по мукам», в мальчишеских «Судьбе барабанщика» и «Белеет парус одинокий», в плутовских романах Ильфа и Петрова и такой же «Судьбе барабанщика». Может показаться странным, но уровень частотности *она* для «Блистающего мира» намного выше, чем для «Бегущей по волнам».

в) Высокая частотность *как*, которую соблазнительно соотнести с обилием сравнительных оборотов, характерна для метафорической «Зависти»

⁵ В строках Мандельштама «...вослед за Персефоной, / Слепая ласточка бросается к ногам / С стигийской нежностью и веткою зеленой. // Навстречу беженке спешит толпа теней, / Товарку новую встречая причитаньем, / И руки слабые ломают перед ней / С недоумением и робким упованием. // Кто держит зеркальце, кто баночку духов, — / Душа ведь женщина, ей нравятся безделки...» — сплошной ряд из 14 существительных женского рода не прерван ни одним существительным мужского рода, и, наоборот, в строках «И снова, паровозными свистками / Разорванный, скрипичный воздух слит. // Огромный парк. Вокзала шар стеклянный. / Железный мир опять заморожен. / На звучный пир в элизиум туманный / Торжественно уносится вагон: / Павлиний крик и рокот фортепьянный. / Я опоздал. Мне страшно. Это — сон. // И я вхожу в стеклянный лес вокзала, / Скрипичный строй в смятении и слезах. / Ночного хора дикое начало / И запах...» — ряд из 15 существительных мужского рода не прерван ни одной формой женского: такие случаи редки, но возможны.

Олеша, трех поздних и двух «немецких», но не для трех ранних «русских» романов Набокова, для «Смерти Вазир-Мухтара», обеих вещей Федина, обеих вещей Шмелева и, что примечательно, для каждого текста Грина.

г) Обилие *ты* маркирует, как кажется, тексты с двумя параметрами: большой ролью диалога и неформальным, неофициальным характером этого диалога (Гайдар, Леонов, Платонов), пониженный уровень означает отказ от одного из этих признаков и приводит к пониженной роли *ты*: Вагинов, Зощенко, Пильняк, Тынянов, Ильф и Петров, ряд романов Набокова.

Заключение

Проведенный анализ наиболее частотной лексики в русских романах межвоенного периода, с одной стороны, проливает свет на общие лексические и отчасти на задаваемые лексикой грамматические черты текстов, позволяющие говорить о некотором сходстве текстов, выраженном на количественном уровне; а с другой стороны, помогает зафиксировать индивидуальные повествовательные, коммуникативные и смысловые особенности романов, а также специфику построения системы персонажей тех или иных текстов. При этом анализ лишь 25 наиболее частотных слов не вступает в противоречия с данными Delta. То, что верхняя часть списка показывает те же тенденции, что массивы в 100, 200 и 500 словоформ, само по себе почти наверняка предсказуемо. Естественно начинать анализ работы алгоритма с области, обнимающей наибольшее количество случаев. Согласно гипотезе, из которой мы исходим, и на более низких «этажах» частотного списка такими же разнообразными и прихотливыми способами сочетаются те же тенденции: идиосинкразические речевые привычки (идиолекты) авторов. Конкуренция языковых средств с близкими функциями и семантикой, связь служебных, местоименных и самых частых именных, глагольных и наречных слов с синтаксисом, повествовательной структурой текста и семантикой.

Мы попытались раскрыть некоторые механизмы кластеризации стиля по частотным словам: сколько слов из верхней части списка совпадает по уровню частотности у разных авторов, каково соотношение нейтральной и маркированной частотности, какова конкуренция близких по употреблению слов, каковы ресурсы связи местоименных и служебных слов с семантикой, композицией и жанром текста. В дальнейшем нужно расширить зону анализа, выявить менее очевидные соответствия и более подробно проследить «анатомию» парадоксальных случаев кластеризации.

Литература

- Андреев В. С. Моделирование индивидуального стиля американских поэтов-романтиков // Вестник Тамбовского университета. Серия: Гуманитарные науки. 2012. № 10. С. 47–52.
- Андреев В. С., Павлова Л. В., Романова И. В. Поэтика цвета «*cattleya labiata*» в лирике Владимира Набокова // Новый филологический вестник. 2021. № 3 (58). С. 214–224.
- Артемova О. Г. Языковые ключи к английской литературе от Шекспира до Фаулза. Воронеж: Наука-ЮНИПРЕСС, 2020. 596 с.
- Введенский А. М. Частотность предлога «на» в древнерусских текстах // Древняя Русь. Вопросы медиевистики. 2013. № 3 (53). С. 26–27.
- Двинятин Ф. Н., Ковалев Б. В. «Сирин не уступает Леонову»: метод Delta для стилеметрического анализа русских романов межвоенного периода // Шаги / Steps. Т. 10. № 3. 2024. С. 207–229.
- Зевахина Т. С., Филиппова М. М. Об одном маркемологическом исследовании художественных текстов, или Можно ли свести четыре века английской литературы к двум ключевым словам // Вестн. Моск. ун-та. Серия 9. Филология. 2024. № 1. С. 111–124.
- Кашкина А. В. Сравнительный маркемный анализ: проблемы и перспективы // Язык, коммуникация и социальная среда: сб. науч. трудов. Воронеж: Воронеж. гос. ун-т; Изд. дом Алейниковых, 2010. Вып. 8. С. 228–235.
- Кретов А. А., Катov М. В. Сквозь призму маркем: Н. В. Гоголь в ближайшем контексте русской литературы // Вестник ВГУ. Серия: Лингвистика и межкультурная коммуникация. 2009. № 2. С. 12–21.
- Кретов А. А., Фаустов А. А. Понятие маркемы и предварительные итоги маркемного анализа русской литературы // ВГУ. Серия: Лингвистика и межкультурная коммуникация. 2017. № 4. С. 16–31.
- 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.
- Национальный корпус русского языка [Электронный ресурс]. URL: <https://ruscorpora.ru> (дата обращения: 06.12.2024).
- Орехов Б. В. Текст и перевод Владимира Набокова через призму стилеметрии // Новый филологический вестник. 2021. № 3 (58). С. 200–213.
- Burrows J. Delta: a measure of stylistic difference and a guide to likely authorship // Literary and Linguistic Computing. 2002. № 3 (17). P. 267–287.
- Calvo Tello J. Delta inside Valle-Inclán: stylometric classification of periods and groups of his novels // Romanische Studien. 2019. № 6. P. 151–163.
- Eder M. Style-Markers in authorship attribution. A cross-language study of the authorial fingerprint // Studies in Polish Linguistics. 2011. № 6. P. 99–114.
- Eder M., Rybicki J., Kestemont M. Stylometry with R: a package for computational text analysis // R Journal. 2016. № 1 (8). P. 107–121.

- Lutoslawski W. The origin and growth of Plato's logic. London: Longmans, 1897. 548 p.
- Michéa R. Mots fréquents et mots disponibles. Un aspect nouveau de la statistique du langage // Les langues modernes. 1953. № 47. P. 338–344.
- Mosteller F., Wallace D. L. Inference in an Authorship Problem // Journal of the American Statistical Association. 1963. Vol. 58. № 302. P. 275–309.
- Rybicki J., Eder M. Deeper Delta across genres and languages: do we really need the most frequent words? // Literary and Linguistic Computing. 2011. Vol. 26. № 3. P. 315–321.

References

- Andreev V. S. [Modeling of Individual Styles of American Romantic Poets]. *Vestnik Tambovskogo universiteta. Seriya: Gumanitarnye nauki*, 2012, no. 10, pp. 47–52. (In Russ.)
- Andreev V. S., Pavlova L. V., Romanova I. V. [The Poetics of Color “Cattleya Labiata” in the Lyrics by Vladimir Nabokov]. *Novyi filologicheskii vestnik*, 2021, no. 3, vol. 58, pp. 214–224. (In Russ.)
- Artemova O. G. *Yazykovye klyuchi k angliiskoi literature ot Shekspira do Faulza* [Language keys to the English literature from Shakespeare to Fowles]. Voronezh, NAUKA-UNIPRESS Publ., 2020. 596 p.
- Burrows J. Delta: a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 2002, no. 3, vol. 17, pp. 267–287. (In Eng.)
- Calvo Tello J. Delta inside Valle-Inclán: stylometric classification of periods and groups of his novels. *Romanische Studien*, 2019, no. 6, pp. 151–163. (In Eng.)
- Dviniatin F. N., Kovalev B. V. [“Sirin is not inferior to Leonov”: Burrows's Delta for stylometric analysis of Russian novels of the interwar period]. *Shagi / Steps*, 2024, no. 3, vol. 10, pp. 207–229. (In Russ.)
- Eder M. Style-Markers in authorship attribution. A cross-language study of the authorial fingerprint. *Studies in Polish Linguistics*, 2011, no. 6, pp. 99–114. (In Eng.)
- Eder M., Rybicki J., Kestemont M. Stylometry with R: a package for computational text analysis. *R Journal*, 2016, no. 1, vol. 8, pp. 107–121. (In Eng.)
- Kashkina A. V. [Comparative markeme analysis: challenges and perspectives]. *Yazyk, kommunikaciya i social'naya sreda* [Language, communication and social environment]. Voronezh, Voronezh State University Publ., 2010, pp. 228–235. (In Russ.)
- Kretov A. A., Katov M. V. [Through the prism of markeme: N. V. Gogol in the immediate context of Russian literature]. *Vestnik Voronezhskogo gos. un-ta. Ser. Lingvistika i mezhkulturnaya kommunikatsiya*, 2009, no. 2, pp. 12–21. (In Russ.)
- Kretov A. A., Faustov A. A. [The Notion of markeme and preliminary findings of markeme analysis of Russian literature]. *Vestnik Voronezhskogo gos. un-ta. Ser. Lingvistika i mezhkulturnaya kommunikatsiya*, 2017, no. 4, pp. 16–31. (In Russ.)
- Lutoslawski W. *The origin and growth of Plato's logic*. London, Longmans Publ., 1897. 548 p.

- Michéa R. Mots fréquents et mots disponibles. *Un aspect nouveau de la statistique du langage* Les langues modernes, 1953, no. 47, pp. 338–344. (In French)
- Mikheev M. Yu., Erlikh L. I. [Idiostyle profile and determination of text authorship by frequencies of function words]. *Nauchno-tekhnicheskaya informaciya. Seriya 2: Informacionnye processy i sistemy*, 2018, no. 2, pp. 25–34. (In Russ.)
- Mosteller F., Wallace D. L. Inference in an Authorship Problem. *Journal of the American Statistical Association*, 1963, no. 302, vol. 58, pp. 275–309. (In Eng.)
- Natsional'nyi korpus russkogo yazyka* [Russian National Corpus]. Available at: <http://ruscorpora.ru/> (accessed 06.12.2024).
- Orehhov B. V. [Text and translation by Vladimir Nabokov through the prism of stylometry]. *Novyi filologicheskii vestnik*, 2021, no. 3, vol. 58, pp. 200–213. (In Russ.)
- Rybicki J., Eder M. Deeper Delta across genres and languages: do we really need the most frequent words? *Literary and Linguistic Computing*, 2011, no. 3, vol. 26, pp. 315–321. (In Eng.)
- Vvedenskii A. M. [Frequency of the preposition “na” in Old Russian texts]. *Drevnyaya Rus'. Voprosy medievistiki*, 2013, no. 3, vol. 53, pp. 26–27. (In Russ.)
- Zevakhina T. S., Filippova M. M. [On One Markemological Study, or Can Four Centuries of English Literature Be Reduced to Two Keywords]. *Lomonosov Philology Journal. Series 9. Philology*, 2024, no. 1, pp. 111–124. (In Russ.)