АВТОМАТИЧЕСКАЯ ПЕРИОДИЗАЦИЯ АВТОРСКИХ КОРПУСОВ

Балуева Д. В. (baluevadaria@ya.ru)
РГГУ, Москва

This paper propose a method of automatic periodization of author's oeuvre. We worked under the assumption that the author's style changes over time. So it should be possible to divide all author's texts on at least two most unlike groups taking into account the chronology (author's early and late oeuvre) The method is applied to 73 single-author corpora taken from the Poetic subcorpus of the Russian National Corpus, all texts are annotated for date. We splitted temporally ordered corpora in two parts in all possible ways and compared received parts using the cosine distance and character 4-grams. Splits with the maximal distance were considered as variants of periodization and checked. It turns out that our method allows to divide author's oeuvre into two most unlike groups. Several results are consistent with history of literature. This method may be improved and used for studying evolution of individual style: for authors presented in the RNC as well as for others.

Keywords: authors' corpora, automatic periodization, Russian poetry, Russian National Corpus, text similarity.

Введение

Цель нашей работы – предложить и проверить метод автоматической периодизации творчества. Мы исходим из предположения о том, что авторский стиль изменяется со временем. Тогда, если упорядочить все тексты какого-либо автора по дате создания, то внутри них можно провести одну или несколько границ, разделяющих тексты на максимально непохожие между собой группы. Такие группы можно будет условно назвать ранним и поздним творчеством (если их две), ранним, «зрелым и поздним творчеством (если их три) и т. п.

Подобные исследования уже проводились ранее, например, автоматической периодизации корпусов посвящена работа [Alsudais, Tchalian 2016]. Задача исследования – выделить тематические периоды в корпусе текстов новостей. Работы [Reeve 2018; Salgaro, Rebora 2018] посвящены периодизации художественной прозы. Авторы проверяют возможность статистически отделить ранние и поздние произведения. Периодизация творчества с помощью точных методов – одна из задач стилометрии, компьютерной стилистики. [Neal 2017; Скоринкин 2018: 51]

Периодизацию корпуса мы понимаем как разделение текстов, размеченных и отсортированных по дате создания, на максимально непохожие между собой группы с сохранением хронологии. Основа периодизации – сравнение текстов, написанных раньше, с текстами, написанными позже. В компьютерной лингвистике для сравнения текстов используются различные меры расстояния [Rayson et al. 2000; Burrows 2002; Gomaa, Fahmy 2013] Сравнения могут проводиться на разных языковых уровнях: лексики [Holmes 1994], служебных слов [Burrows 2002], символьных n-грамм [Kjell et al.1994] и других. Меры неоднократно уточнялись и оценивались, например в работах [Kilgariff 2001, Hoover 2004, Argamon 2008, Eder 2015].

Материал работы

Материалом нашей работы послужили тексты из поэтического подкорпуса Национального корпуса русского языка с размеченным авторством и датой создания. Мы используем корпуса 73 авторов, объемом 50 тысяч токенов и более. При подсчете токенов мы учитывали все последовательности из букв, цифр и дефисов. Полный список авторов, использованных в работе, приводится в приложении. (Таблица 2)

Все тексты в поэтическом подкорпусе размечены по дате создания. Даты могут быть указаны по-разному:

1) С точностью до дня создания: 1965.08.05
2) Только год: 1903
3) Период создания длиной в несколько лет: 1903-1905

В качестве единицы времени для периодизации мы выбрали один год. Первый тип дат мы сокращали до года, так как текстов с конкретным днем создания очень мало. Сложность для нашего исследования представляют тексты, размеченные по третьему принципу. С одной стороны, такая разметка может отражать фактические годы работы над произведением. Например, годы создания “Поэмы без героя” Анны Ахматовой, над которой она действительно работала более двадцати лет, в корпусе указаны как “1940- 1965”. С другой стороны, дата создания некоторых текстов может быть указана примерно, когда подлинная неизвестна. Произведения с неточной датировкой затрудняют периодизацию творчества, поэтому мы исключили такие тексты из рассмотрения.

Периодизация корпусов

Тексты каждого автора группировались по году создания и сортировались по возрастанию года. Затем, соблюдая хронологию, мы разбивали тексты на две части всеми возможными способами. Каждый раз получившиеся группы сравнивались. Тексты мы сравнивали, вычисляя косинусное расстояние между частотными векторами символьных *n*-грамм.

Косинусная мера устойчива к изменению длин векторов, то есть результат не зависит от длины сравниваемых текстов. [Evert et al. 2017] Поэтому она подходит для сравнения текстов стихотворений, которые гораздо короче прозаических, а также текстов, написанных за неравные промежутки времени (например, за 10 и 50 лет).

Символьные *n*-граммы признаются одним из самых эффективных признаков для определения авторства [Juola 2006; Houvardas, Stamatatos 2006]. Они могут говорить «обо всем по-немногу»: содержать информацию о лексике (корни слов или служебные слова полностью), морфологии (морфемы) [Kestemont 2014: 60-62], фонологии [Piperski 2019: 8] и структуре текста (пунктуация и деление на строки и абзацы).

Мы привели все тексты к нижнему регистру и разбили на символьные *n*-граммы длины четыре. Их эффективность, по сравнению с *n*-граммами длины 2,3 обнаружилась в исследовании [Piperski 2019], посвященном определению авторства на материале, аналогичном материалу данной работы.

Покажем, например, периодизацию корпуса поэта Бориса Корнилова. В корпусе собраны стихотворения за 12 лет: с 1925 по 1936 год. Сначала мы сравнили все тексты за 1925 год со всеми текстами за 1926-1936 годы. Затем тексты за 1925-1926 годы с

текстами за 1927-1936 годы. И так далее, пока не дошли до сравнения текстов за 1925-1935 годы с текстами за 1936 год. Всего 11 возможных разбиений и 11 сравнений:

Номер	Период 1	Период 2	Расстояние
1	1925	1926-1936	0.539
2	1925-1926	1927-1936	0.253
3	1925-1927	1928-1936	0.142
4	1925-1928	1929-1936	0.116
5	1925-1929	1930-1936	0.1
6	1925-1930	1931-1936	0.09
7	1925-1931	1932-1936	0.08
8	1925-1932	1933-1936	0.111
9	**1925-1933**	**1934-1936**	**0.136**
10	1925-1934	1935-1936	0.123
11	1925-1935	1936	0.105

Таблица. 1 Поиск локального максимума расстояния

При девятом разбиении корпуса (1925-1933 и 1934-1936) расстояние между частями оказалось больше, чем два соседних (0.136 > 0.11, 0.136 > 0.123). Чем больше расстояние между частями, тем меньше они похожи. Это может значить, что после 1933 года стиль поэта изменился сильнее, чем после 1932 года или после 1934 года.

В корпусах некоторых авторов, как и у Корнилова, нашелся только один локальный максимум, у других не нашлось вообще, а у третьих нашлось несколько. Например, при разбиении корпуса Анны Ахматовой удалось найти 12 локальных максимумов, а при разбиении корпуса Николая Гумилева – ни одного.

Рассматривая разные типы небуквенных сочетаний, мы обратили внимание, что у некоторых авторов очень частотны сочетания из четырех знаков препинания. Встречались редкие эмоциональные сочетания ида ‘?..»’, ‘!..»’ или даже ‘????’, которые безусловно являются значимыми. Но в то же время, в некоторых корпусах было много сочетаний вида ‘……’ (четыре троеточия подряд) или комбинаций из нескольких троеточий и других символов.

Судя по всему, последовательности из троеточий используются, чтобы обозначить пропущенные фрагменты текста. Как, например, здесь:

Его рассматривали как святого,
[. ……..]
[……..]
[……..]

Как [. ……. .] ядом,

[М. А. Тарловский. «Охотясь на пещерного медведя...» (1951.06.29)]

Мы проверили, будет ли значимо ли исключение таких сочетаний. Сочетания, содержащие более одного троеточия, встретились в текстах 57 авторов из 73. Удаление отразилось на результате: например, больше всего троеточий было в корпусе Анны Ахматовой (около 0.2% от всех сочетаний), и локальных максимумов стало 8 вместо 12. Количество локальных максимумов расстояния для нашего эксперимента – это

количество вариантов периодизации, и чтобы исключить незначимые варианты, мы решили не учитывать n-граммы, включающие больше одного троеточия.

Проверка полученных разбиений

Каждый вариант периодизации мы проверили. Сначала мы дополнительно делили и смешивали тексты разных периодов:

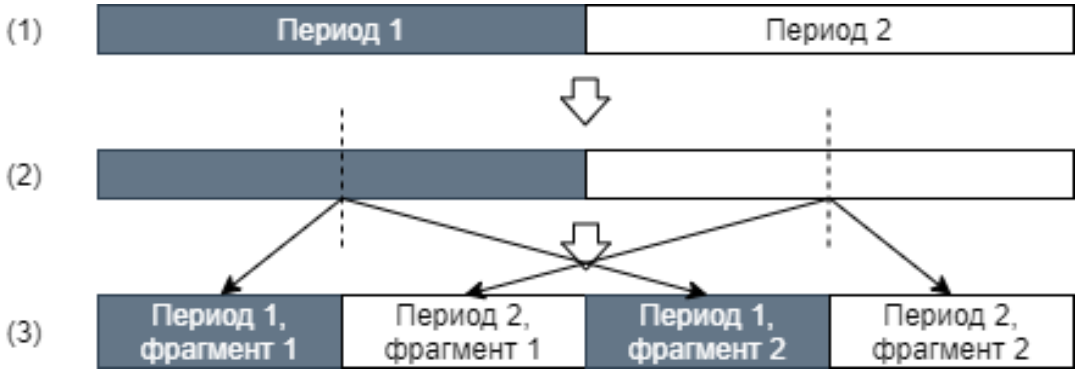

Рис. 1. Дополнительное разбиение и смешивание полученных групп текстов

Каждое полученное на первом этапе разбиение разделило корпус на две части: предположительно, раннее творчество (Период 1) и, предположительно, позднее творчество (Период 2) **(1)**. Каждый из этих периодов мы вручную разбили пополам **(2)**. Полученные фрагменты перемешали и соединили так, чтобы границы разных фрагментов одного периода не накладывались **(3)**.

Затем мы проверяли, получится ли отделить фрагменты разных периодов друг от друга. Для каждого варианта периодизации (локального максимума расстояния, полученного на предыдущем этапе) мы сделали по четыре проверки:

Рис. 2. Схема проверки полученных разбиений

Фрагменты разных периодов мы отделяли так же, как искали границы на предыдущем этапе: делили тексты на две части, сравнивали их и искали разбиение с наибольшим расстоянием. Если расстояние было наибольшим именно на границе фрагментов, мы считали проверку успешной.

Четыре раза нам удалось подтвердить только периодизации корпусов А. К. Толстого, И. Г. Эренбурга и М. А. Волошина. В остальных случаях удачными оказывались 3/4, 2/4, 1/4 или 0 проверок.

Равноценны ли все проверки? В 1, 3 и 4 проверках (см. Рис. 2) мы сравниваем фрагменты, которые хронологически отделены друг от друга еще одним промежутком

времени. А во второй проверке наоборот сравниваются наиболее близкие фрагменты: фактически, мы снова пытаемся провести границу в том же месте.

Судя по всему, разбиения, подтвержденные четыре раза, отражают самую явную временную границу в творчестве поэтов. Корпуса делятся на две наиболее непохожих между собой группы текстов. При этом фрагменты одного периода похожи друг на друга, но не похожи на фрагменты другого периода.

Можно предположить, что для разбиений, подтвержденных три раза из четырех, более важными являются проверки 1, 3 и 4. Если наш метод действительно позволяет делить весь корпус на две максимально непохожие между собой группы, то важнее, чтобы друг от друга лучше отделялись более удаленные временные промежутки.

В следующем разделе мы сравним периодизации некоторых корпусов с тем, что уже известно о творчестве авторов.

Литературоведческая интерпретация результатов

1) М. А. Волошин: 1899-1916 и 1917-1931 годы.

Данную периодизацию удалось подтвердить четыре раза, и интересно, что она согласуется с тем, что о своем творчестве писал сам Волошин. Максимилиан Волошин указал, в каком порядке и какими сборниками должны быть выпущены его произведения. Произведения, написанные с 1915 года он выделил в отдельную книгу *«о войне и революции».* [Филиппов 2005] Фактически, так Волошин сам обозначил поздний тематический период своего творчества. Наша периодизация отличается от периодизации Волошина всего на несколько лет.

2) И. Г. Эренбург: 1910-1924 и 1939-1958 годы.

Данное разбиение тоже удалось проверить четыре раза. 15 лет между 1924 и 1939 годом Эренбург действительно ничего не писал. [Фрезенский 1999] Фрезенский называет творчество Эренбурга с 1939 года поздним периодом.

3) А. С. Пушкин: первый вариант – 1813-1829 и 1830-1836 годы.

Периодизация творчества А. С. Пушкина – интересный пример для нашего исследования. При разбиении его корпуса нашлось два локальных максимума расстояния, и оба эти варианта подтвердились три раза.

В статье [Фомичев 1982] приводится периодизация, называемая биографической: «Лицей – Петербург – Юг – Михайловское – После 1825 года – Последние (т. е. 30-е) годы». Эта схема опирается на важнейшие события из жизни поэта. Аргумент в пользу биографической периодизации творчества Пушкина – то, что *«он по преимуществу поэт-лирик, и потому резкие повороты его судьбы чутко отражались в его поэзии»*.

Осень 1830 года, когда Пушкин во время эпидемии холеры находился на карантине в Болдино, называется вершиной творчества поэта: было написано множество произведений, закончен роман «Евгений Онегин». Выделение позднего периода (с 1830 года) связано и с большими изменениями авторского стиля.

По нашей автоматической периодизации граница в творчестве Пушкина проходит тоже между 1829 и 1830 годами. Успешными оказались все проверки, кроме второй: когда сравнивались тексты за 1830-1833 и 1822-1829 годы. Как было сказано в предыдущем разделе работы, результаты второй проверки наименее показательны для

нашего метода периодизации. Можно считать, что данный вариант лучше всего отражает разбиение творчества Пушкина на два периода, но мы получили еще один.

Второй вариант – 1813-1821 и 1822-1836

Неуспешной оказалась проверка 4: не удалось отделить друг от друга конец первого периода (1818-1821) и конец второго (1830-1836). Эти группы текстов отделены друг от друга промежутком в 9 лет, но различие между ними оказалось незначительным.

Однако Фомичев пишет, что некоторые исследователи выделяют и различия в лирике Пушкина 1817-1821 и 1821-1822 годов. Эта периодизация охватывает только небольшой временной промежуток, и мы провели границу между 1921 и 1822 годами, но не разделили корпус на две части.

4) И. А. Бунин: 1888-1902 и 1903-1952

Для этого разбиения удались только две проверки (№ 2 и №4), то есть, мы смогли однозначно отделить друг от друга наиболее хронологически близкие (1903-1918 и 1896-1902) и наиболее далекие фрагменты. (1888-1895 и 1919-1952) Значит, самые ранние и самые поздние произведения Бунина действительно не похожи между собой, и изменения заметны сразу после 1902 года.

В статье [Балановский 2011] приводится обзор периодизаций творчества Бунина. Одна из самых первых – версия С. Родзевича, предложенная еще при жизни Бунина, в 1914 году [Родзевич 1914]. Родзевич уже тогда разделил раннюю поэзию Бунина на два периода: «конец 1880-х – 1902 г. и 1903 – 1906 гг». Так как это одна из самых первых существующих периодизаций творчества Бунина, последующие исследования дополняют и уточняют ее. Наша автоматическая граница оказалась в том же месте, где ее проводил литературовед начала 20 века.

Заключение

Мы предложили и проверили метод автоматической периодизации авторских корпусов. Если вариантов (локальных максимумов расстояния) обнаружить не удалось, это может свидетельствовать как о том, что стиль поэтов незначительно менялся с течением жизни. При этом если вариант не удавалось подтвердить полностью, это может значить, что реальная периодизация творчества должна быть более дробной, чем разделение на две части.

Тем не менее у нашей работы есть несколько ограничений. Во-первых, неточными могут быть и литературоведческие периодизации, с которыми мы сравнивали свои результаты. Во-вторых, литературоведы учитывали жанровое и тематическое многообразие творчества, биографические факты, проводили стиховедческий анализ. Мы же сравнивали тексты только на языковом уровне. Один из вариантов продолжения нашего исследования – подбор признаков, наиболее релевантных для периодизации: отдельных видов *n*-грамм (например, небуквенных) и неязыковых признаков (метр, строфика, жанр и т.д.). Для последнего может быть использована богатая метаразметка поэтического подкорпуса НКРЯ [Гришина и др. 2009].

Отметим также, что мы с самого начала исключили из рассмотрения тексты, размеченные не одним годом создания, и это могло сильно повлиять на точность периодизации. В будущих исследованиях могут быть предложены способы учета всех текстов.

Библиография

Электронные ресурсы

1. Поэтический подкорпус Национального корпуса русского языка http://www.ruscorpora.ru/new/search-poetic.html
2. Список всех авторов, представленных в поэтическом подкорпусе http://www.ruscorpora.ru/new/poet-list.html

Литература

1. Балановский Р. М. Проблема периодизации творчества И. А. Бунина в русской критике и литературоведении. Вестник Череповецкого государственного университета. Т. 1. Выпуск 1 (28). Череповец, 2011.
2. Гришина Е. А., Корчагин К. М., Плунгян В. А., Сичинава Д. В. Поэтический корпус в рамках НКРЯ: общая структура и перспективы использования // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009.
3. Родзевич С. В поисках Атлантиды: (Поэзия И. А. Бунина) Тип. А. М. Пономарева п. у. И. И. Врублевского, Киев, 1914.
4. Скоринкин Д. А. Семантическая разметка художественных текстов для количественных исследований в филологии (на примере романа "Война и мир" Л.Н. Толстого): дис. канд. филол. наук. М., 2018.
5. Филиппов Г. В. Волошин Максимилиан Александрович. Русская литература XX века. Прозаики, поэты, драматурги: библиогр. словарь : в 3 т. под ред. Н. Н. Скатова. М. ОЛМА-ПРЕСС Инвест. Т. 1. А-Ж. с. 419-423. 2005.
6. Фомичев С. А. Периодизация творчества Пушкина: (К постановке проблемы) // Пушкин: Исследования и материалы / АН СССР. Ин-т рус. лит. (Пушкин. Дом). Л.: Наука. Ленингр. отд-ние, 1982.
7. Фрезенский Б. И. Запретный Эренбург. Арион. № 3(23). с. 32–43. 1999.
8. Alsudais, A., & Tchalian, H. (2016). Corpus Periodization Framework to Periodize a Temporally Ordered Text Corpus. *AMCIS*.
9. Argamon S. (2008). Interpreting Burrows’ delta: Geometric and probabilistic foundations. Literary and Linguistic Computing , 23(2): 131–47.
10. Burrows J. (2002). ‘Delta’: a measure of stylistic difference and a guide to likely authorship. Literary and linguistic computing, 17(3), 267-287.
11. Dobson James E. (2019) Critical Digital Humanities: The Search for a Methodology. University of Illinois Press.
12. Eder M. (2015). Taking stylometry to the limits: Benchmark study on 5,281 texts from Patrologia Latina. In Digital Humanities 2015: Conference Abstracts.
13. Evert S., Proisl, T., Jannidis F., Reger I., Pielström S.,Schöch C., & Vitt, T. (2017). Understanding and explaining Delta measures for authorship attribution, Digital Scholarship in the Humanities, Volume 32, Issue suppl_2, December 2017, Pages ii4–ii16.
14. Gomaa W. H., Fahmy A. (2013). A Survey of Text Similarity Approaches International / Journal of Computer Applications, 68(13), April, pp. 13–18.
15. Holmes D.I. (1994). Authorship attribution. Computers and the Humanities, 28(2), 87–106.

16. Hoover D. (2004). Delta prime? Literary and Linguistic Computing , 19(4): 477–95.
17. Houvardas J., & Stamatatos E. (2006). N-Gram Feature Selection for Authorship
18. 11. Identification. In J. Euzenat & J. Domingue (Eds.), Artificial Intelligence: Methodology,
19. Juola P. (2006). Authorship attribution. Foundations and Trends in Information Retrieval , 1(3): 233–334.
20. Kestemont, M. (2014). Function Words in Authorship Attribution. From Black Magic to Theory? Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL), 59–66. https://doi.org/10.3115/v1/W14-0908
21. Kilgarriff A. (2001). Comparing corpora. International Journal of Corpus Linguistics, 6:1, 97–133.
22. Kjell B., Woods W.A., & Frieder O. (1994). Discrimination of authorship using visualization. Information Processing & Management, 30(1), 141–150.
23. Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., & Woodard, D. (2017) Surveying Stylometry Techniques and Applications. ACM Comput. Surv., 50(6), 86:1–86:36.
24. Piperski A. (2019). Authorship Attribution with a Very Naïve Bayes Model and What It Can Tell Us about Russian Poetry. In Computational Linguistics and Intellectual Technologies. Papers from the Annual International. Conference “Dialogue”. Issue 18
25. Rayson P., & Garside R. (2000). Comparing corpora using frequency profiling. / In Proceedings of the Comparing Corpora Workshop at ACL 2000. Hong Kong.
26. Reeve J. P. (2018). “Does ‘Late Style’ Exist? New Stylometric Approaches to Variation in Single-Author Corpora”, in DH2018 Book of Abstracts, ADHO, Mexico City, pp. 478-481.
27. Salgaro M., & Rebora S. (2018). Is “Late Style” measurable? A stylometric analysis of Johann Wolfgang Goethe’s, Robert Musil’s, and Franz Kafka’s late works. Elephant&Castle, Systems, and Applications (pp. 77–86). Berlin, Heidelberg: Springer.

Приложение

Поэт	Количество текстов	Объем корпуса
В. А. Жуковский	695	245119
А. С. Пушкин	904	193440
В. Я. Брюсов	1683	184500
Н. А. Некрасов	446	170401
М. И. Цветаева	1468	166270
В. В. Маяковский	634	133591
М. Ю. Лермонтов	474	132281
Б. А. Слуцкий	1287	123602
В. И. Иванов	1180	123389
А. К. Толстой	301	118361
И. В. Елагин	866	117081
Саша Черный	705	114723
А. Т. Твардовский	346	113028
А. А. Блок	1350	112945
А. Н. Майков	561	112831
Е. А. Евтушенко	345	110066
К. Д. Бальмонт	1004	107136
Б. А. Ахмадулина	394	104068
В. А. Соснора	539	95026
К. К. Случевский	735	94954
Г. Р. Державин	415	93335
Д. Л. Андреев	452	91684
М. А. Кузмин	837	91268
Н. А. Клюев	572	91102
И. Л. Сельвинский	458	83421
Д. С. Мережковский	338	83348
К. М. Симонов	282	81847
И. А. Бунин	767	79893
А. А. Фет	920	79834
Б. Л. Пастернак	531	77599
А. П. Сумароков	283	76353
С. М. Соловьев	430	74714
Д. Самойлов	900	73064
И. Г. Эренбург	710	70186
В. К. Тредиаковский	162	68239
П. А. Вяземский	339	67200
Н. С. Гумилев	523	67091

Поэт	Количество текстов	Объем корпуса
С. В. Петров	608	65882
П. Г. Антокольский	370	65668
Л. Н. Мартынов	674	65005
Я. В. Смеляков	371	64438
В. В. Набоков	591	62880
П. Н. Васильев	244	62596
А. И. Несмелов	451	62445
И. С. Никитин	231	61985
Н. Н. Асеев	375	61667
С. И. Кирсанов	440	61595
Н. М. Языков	358	60001
И. П. Мятлев	114	59303
О. Э. Мандельштам	679	59275
С. А. Есенин	427	58120
С. Я. Надсон	490	57285
Г. Н. Оболдуев	387	57213
И. Северянин	637	56626
М. М. Херасков	64	56615
Н. Байтов	394	56525
А. А. Ахматова	945	56372
З. Н. Гиппиус	474	55471
И. И. Дмитриев	391	55449
О. Ф. Берггольц	392	55251
А. Н. Апухтин	354	54943
Э. Г. Багрицкий	202	54233
Андрей Белый	503	53810
В. В. Хлебников	290	53767
М. В. Ломоносов	145	53230
Б. Ю. Поплавский	547	52981
В. А. Луговской	150	52917
В. И. Майков	112	52259
Б. П. Корнилов	188	50866
М. А. Волошин	299	50834
М. А. Тарловский	315	50735
В. Г. Бенедиктов	271	50329
П. Ф. Якубович	366	50263
Всего:	**38685**	**6081824**

Таблица 2.Авторы, представленные в Поэтическом подкорпусе Национального корпуса русского языка более чем 50000 токенами