Меры расстояния для определения авторства древнегреческих текстов[1]

Аннотация

Хотя классическая филология была одной из первых гуманитарных дисциплин, обратившейся к количественным исследованиям, стилометрия никогда не рассматривалась филологами как самодостаточный метод. Стилистические и языковые особенности — лишь один, не главный и не единственный, инструмент исследователя, а из этих особенностей лишь некоторые могут быть описаны количественно. Более того, количественные методы пригодны только для отрицательного решения вопроса о подлинности, в то время как несомненных оснований для утвердительного решения вообще не существует. В настоящем исследовании сравниваются методы, в англоязычной литературе известные как *distance-based approaches*, то есть подходы, основанные на векторизации текстов через частотности токенов и измерении расстояний между векторами. Оценивается эффективность подобных методов в зависимости от длины отрывка и количества предикторов. Для сравнения привлекается корпус древнегреческой прозы объемом 694 тыс. слов. Наиболее эффективные метрики затем привлекаются для классификации трех спорных текстов.

1. Постановка вопроса

Хотя классическая филология была одной из первых гуманитарных дисциплин, обратившейся к количественным исследованиям, стилометрия никогда не рассматривалась филологами как самодостаточный метод. Суждение об авторстве текста, как писал Фридрих Бласс, должно учитывать данные рукописной традиции, свидетельства современников, соответствие бытовых и исторических реалий времени жизни автора, а также соответствие идей, тем и жанров тому, что известно по подлинным сочинениям автора [Blass, 1892. S. 289–295 = Бласс, 2016. С. 186–194]. Стилистические и языковые особенности — лишь один, не главный и не единственный, инструмент исследователя, а из этих особенностей лишь некоторые могут быть описаны количественно. В этом смысле и современные методы «атрибуции

[1] Я благодарю Бориса Орехова и Даниила Скоринкина за замечания к первым версиям этой статьи, позволившие существенно доработать как сам текст, так и код. Я также признательна Артему Юнусову, который обратил мое внимание на две работы Р. Ойкена о служебных словах у Аристотеля. — ОА.

авторства», опирающиеся на статистику и технологии, могут играть только вспомогательную роль и не представляют угрозы традиционным подходам [Holmes, 1994. P. 104]. Более того, эти методы (впрочем, как и остальные) «пригодны только для отрицательного решения вопроса о подлинности», а несомненных оснований для утвердительного решения «вообще не существует» [Blass, 1892. S. 292 = Бласс, 2016. С. 189]. Говорят, что Пауль Фридлендер как-то возразил Виламовицу: «Невозможно доказать, что Платон написал “Пир”» [Chambers, 1996. P. 220][2]. По умолчанию традиция достоверна, а бремя доказательства лежит на том, кто ее отвергает: *quivis praesumitur genuinus liber, donec demonstretur contrarium* [Boeckh, 1886. S. 240][3].

Но даже для отвержения авторства стилометрия может быть несовершенным инструментом. Кеннет Довер заметил, что длина предложения в греческом тексте больше говорит о редакторе, чем об авторе [Dover, 1968. P. 108]; как издатель Аристофана, он прекрасно понимал, что наши источники —результат множества решений, причем не только авторских. Требование избавиться от всех посторонних вмешательств понятно [Rudman, 2005], но применительно к древним текстам едва ли выполнимо[4]. С другой стороны, сам автор способен менять свой стиль в зависимости от темы, жанра и драматических особенностей произведения, и подобные стилизации неизбежно ставят под вопрос осмысленность наших экспериментов. «Менексен» почти во всех количественных исследованиях — и новых, и старых — не похож на остальные диалоги Платона [Koentges, 2020], но это мало смущает тех, кто считает диалог подлинным (у Платона нет других сочинений в жанре надгробной речи). Стилистическая близость «Послезакония» к поздним диалогам Платона [Ledger, 1989. P. 150] никого не убеждает: автором этого небольшого сочинения по-прежнему признается Филипп Опунтский [Tarán, 1975].

Доверять или не доверять количественным данным — решает исследователь. Древние авторы не могут, как Джоан Роулинг, подтвердить или опровергнуть результаты автоматической классификации текста [Juola, 2013]; а если бы и могли, нам, возможно, не стоило бы их слушать: представления об авторстве меняются со временем [Пешков, 2016]. Феномен «школьной аккумуляции» в античности указывает на то, что «автор» и «не автор» — это, возможно, не бинарная переменная, а лишь два

[2] О важных импликациях этого тезиса применительно к античной эпистолографии см.: [Forcignanò, Martinelli Tempesta, 2023].
[3] «Всякая книга признается подлинной, пока не доказано обратное» (лат.).
[4] О проблеме «неустойчивости» античных и средневековых текстов и подходах к их реконструкции см.: [Шумилин, 2020. С. 33 с прим. 11 *et passim*].

значения непрерывной величины, между которыми есть еще «почти автор» и «не вполне автор» [Thesleff, 2009. P. 351; Thesleff, 2023]. Это ставит нас перед более общим вопросом о том, кто или что является носителем «авторской функции» [Фуко, 2008] и что обеспечивает тождественность этого носителя. Если ранние «Европейцы» Генри Джеймса и его же поздние «Послы» не опознаются как сочинения одного автора, считать ли это ошибкой классификации [Hoover, 2004. P. 457]? Как в известной загадке о корабле Тесея, авторский стиль, постепенно меняясь, может обновиться настолько, что перестает быть «тем же». Эта загадка не решается количественными способами.

Любой метод имеет ограниченное применение, но это не значит, что мы должны признать стилометрию видом «шаманизма» [Love, 2002. P. 159]. Впрочем, если бы 80% ударов в бубен достигали своей цели, у нас были бы основания присмотреться и к шаманским практикам — а некоторые способы автоматической классификации текстов позволяют добиться даже большей точности. Лучшему пониманию и возможностей, и ограничений количественных методов способствуют их испытания на известном материале, и именно такое испытание документируется в этой статье. Разумеется, нет недостатка в сравнениях на материале других языков, прежде всего современных европейских [Stanikūnas *et al*., 2017; Eder, 2011; Jannidis, Lauer, 2014; Rybicki, Eder, 2011], восточных [Ahmed, 2019], а также латинском [Kestemont *et al*., 2016; Eder, 2015]. Но универсальных подходов нет, и то, что показывает хорошие результаты на одном корпусе, не обязательно сработает на другом. Эксперименты с древнегреческими текстами до сих пор проводились лишь в довольно ограниченных масштабах [Фоминых, 2017; Алиева, 2022], и это, как представляется, оправдывает наше усилие.

2. Меры расстояния

Мы выбрали группу методов, в англоязычной литературе известных как *distance-based approaches*, то есть «подходы, основанные на расстоянии» [Savoy, 2020. P. 33]. В их основе лежит идея о том, что текст или группа текстов могут быть представлены в виде вектора — упорядоченного множества значений, которые называются координатами или компонентами вектора. Для каждой пары векторов может быть вычислено расстояние или сходство между ними; минимальное расстояние или максимальное сходство будут указывать на возможного автора. Обратим внимание, что функция считается метрикой расстояния, только если она удовлетворяет критериям неотрицательности, идентичности и симметричности, а также отвечает

дополнительному условию — неравенству треугольника [Хачумов, 2012]. В прочих случаях используется понятие «расхождение» [Cha, 2007. P. 300].

Для сравнения мы отобрали несколько функций, уже применявшихся в стилометрических исследованиях и показавших хорошие результаты. Манхэттенское расстояние, оно же расстояние городских кварталов, лежит в основе метода Берроуза [Burrows, 2002]. В качестве альтернативы предлагалось использовать евклидово расстояние [Argamon, 2008], а также косинусное сходство[5] [Smith, Aldridge, 2011. P. 79–80], в том числе с предварительной стандартизацией [Evert, Proisl *et al.*, 2017]. Стандартизация признаков по z-оценке показывает, на сколько стандартных отклонений значения признака больше или меньше среднего арифметического. Высокая точность классификации достигались с использованием сходства Ружечки, оно же *minmax* [Koppel, Winter, 2014; Kestemont *et al.*, 2016]. 1 – *minmax* эквивалентно расстоянию Танимото, и наоборот [Cha, 2007. P. 302]. Канберрское расстояние рекомендовалось для использования на арабском корпусе [Ahmed, 2019], а расстояние Кларка, среди прочих, — на английском и французском [Kocher, Savoy, 2019]. Из семейства энтропийных расстояний мы возьмем расхождение Джеффриса [Деза, Деза, 2008. С. 221], которое представляет собой симметричную версию расхождения Кульбака-Лейблера; последнее называют также относительной энтропией [Savoy, 2020. P. 39–42]. Поскольку перекрёстная энтропия H(P, Q) для распределений P и Q определяется как сумма энтропии H(P) и относительной энтропии $D_{KL}(P, Q)$, отдельно перекрёстную энтропию мы в этой работе не рассматриваем[6]. Кроме того, протестировано расстояние Лаббе [Labbé, Labbé, 2006; Labbé, 2007; Cortelazzo *et al.*, 2013]. Формулы приведены в Табл. 1.

Табл. 1. Меры расстояния и сходства

1	$D_{Manhattan}$	$\sum_{i=1}^{n} \lvert P_i - Q_i \rvert$
2	$D_{Euclidean}$	$\sqrt{\sum_{i=1}^{n} \lvert P_i - Q_i \rvert^2}$

[5] Также называемое подобность Орчини, угловая подобность или нормированное скалярное произведение [Деза, Деза, 2008. С. 264].
[6] Об энтропии в целом см.: [Shannon, 1948]; перекрестная энтропия для классификации текстов: [Juola 1997; Juola, 1998; Juola, Baayen, 2005] относительная энтропия для классификации текстов: [Zhao, Zobel, Vines, 2006] и [Zhao, Zobel, 2007].

3	S_{Cosine}	$\frac{\sum_{i=1}^{n} P_i Q_i}{\sqrt{\sum_{i=1}^{n} P_i^2}\sqrt{\sum_{i=1}^{n} Q_i^2}}$
4	$D_{Tanimoto}$	$\frac{\sum_{i=1}^{n}\left(max(P_i, Q_i) - min(P_i, Q_i)\right)}{\sum_{i=1}^{n} max(P_i, Q_i)}$
5	$D_{Canberra}$	$\sum_{i=1}^{n} \frac{\lvert P_i - Q_i \rvert}{P_i + Q_i}$
6	D_{Clark}	$\sqrt{\sum_{i=1}^{n} \left(\frac{\lvert P_i - Q_i \rvert}{P_i + Q_i}\right)^2}$
7	$D_{Jeffreys}$	$\sum_{i=1}^{n} (P_i - Q_i) ln \frac{P_i}{Q_i}$
8	$D_{Labbé}$	$\frac{\sum_{i=1}^{n} \lvert P_i - Q_i \rvert}{2N_P}$

3. Задачи эксперимента и программные средства

В ходе эксперимента мы ставили перед собой следующие задачи:

1) выяснить, какие меры расстояния дают наибольшую точность на отрывках разной длины с использование разного числа переменных;
2) установить, обнаруживаются ли различия при использовании стандартизированных и нестандартизированных значений частотности (для тех методов, которые это допускают);
3) сравнить точность атрибуции при использовании словоформ или трехсимвольных энграм;
4) сделать выводы о том, на каких текстах классификатор чаще ошибается;

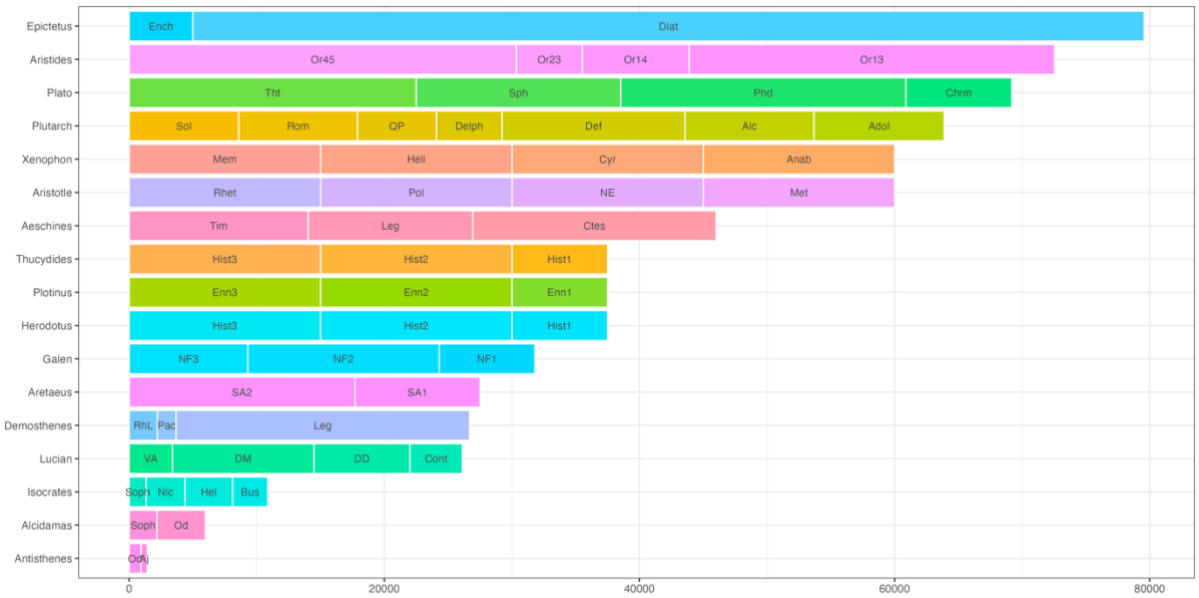
Все вычисления и визуализации выполнены в R, при этом токенизация (деление на слова или энграмы) выполнена с использованием библиотеки Stylo; для составления матриц расстояния или сходства привлекалась библиотека Philentropy, для других вычислений и визуализаций использовались Tidyverse и Tidymodels. Переход от квадратной матрицы расстояния или сходства к классификации осуществлялся при помощи «доморощенных» (в смысле *home-brew*) функций get.pred.min и get.pred.max, для каждого ряда матрицы извлекающих имя столбца с минимальным (для расстояний) или максимальным (для сходства) значением в корпусе. При этом, разумеется, не учитывался столбец с тем же самым произведением[7].

[7] Весь код доступен на GitHub: https://github.com/locusclassicus/compareDist/tree/master

4. Корпус

В корпус, подготовленный для этого эксперимента, вошли сочинения древнегреческих прозаиков (историков, врачей, философов и ораторов) из библиотеки Perseus[8]. Объем корпуса —694 тыс. слов. Корпус является несбалансированным в двух отношениях: разные авторы представлены разным количеством текстов, а сами эти тексты неравномерны по объему. Диспропорцию хорошо видно на Рис. 1. Две декламации Антисфена («Одиссей» и «Аякс») в сумме дают меньше полуторы тысячи слов; максимальное число слов у Эпиктета, но из них лишь около 5000 приходится на «Энхиридион». Аристид представлен четырьмя речами (13 «Панафинейская речь», 14 «Похвала Риму», 23 «Священные речи α'», 45 «К Платону о риторике»), а Платон — четырьмя диалогами («Хармид», «Федон», «Софист», «Теэтет»), как и Лукиан («Харон, или Наблюдатели», «Разговоры мертвых», «Разговоры богов», «Продажа жизней»). Среди сочинений Плутарха — жизнеописания («Ромул», «Солон», «Алкивиад»), «Платоновские вопросы», трактат «О том, как юноше слушать поэтические произведения», а также два диалога: «Об упадке оракулов» и «О "Е" в Дельфах».

Рис. 1. Количество слов и текстов на автора



Аристотель представлен выборками из «Метафизики», «Политики», «Риторики» и «Никомаховой этики»; Ксенофонт — выборками из «Анабасиса», «Меморабилий», «Греческой истории» и «Киропедии». Также среди историков Геродот (три выборки из

[8] Perseus, *Canonical Greek Literature*: https://github.com/PerseusDL/canonical-greekLit. Для извлечения текстов использовались пакеты: *XML* (https://cran.r-project.org/web/packages/XML/index.html) и *RPerseus* (https://github.com/ropensci/rperseus).

«Истории») и Фукидид (три выборки из «Истории»); к философам мы добавили Плотина (три выборки из «Эннеад»). Среди текстов Галена в репозитории библиотеки Perseus нам был доступен единственный трактат «О естественных способностях», из которого также взято три выборки; из сочинений врача II в. Аретея — первая и вторая книги «Этиологии и симптомов острых заболеваний». К этому мы добавили две небольшие речи Алкидаманта («Против софистов» и «Одиссей»), три речи Эсхина («Против Ктесифонта», «Против Тимарха», «О преступном посольстве»), четыре речи Исократа («Бусирис», «Елена», «К Никоклу», «Против софистов») и три речи Демосфена («О преступном посольстве», «О мире», «О свободе родосцев»)[9]. Итого 57 текстов (включая выборки) и 17 авторов.

Тексты не подвергались лемматизации, стеммингу, частеречной или синтаксической разметке[10]. Для анализа они были только разделены на токены: словоформы и трехсимвольные энграмы (с сохранением диакритики).

5. Оценивание

Для сравнения каждый метод опробовался на отрывках 1000–7000 токенов (с шагом 500, всего 13) с разным количеством предикторов (mfw) — от 100 до 1000 (с шагом 100, всего 10). Для каждой длины отрывка и mfw проведено 10 итераций, при этом использовались повторные выборки (с замещением из-за наличия в корпусе очень коротких текстов). Таким образом, для оценки каждого метода выполнено $57 \times 10 \times 13 \times 10 = 74100$ классификаций без стандартизации и столько же со стандартизацией (там, где это допускает метрика). Это было проделано сначала на словоформах, потом на энграмах. Исключение составляет метод Лаббе, не предполагающий отбора mfw, а задействующий все данные об абсолютной частотности слов в отрывке (поэтому здесь всего 7410 классификаций). Точность рассчитывалась как количество верных классификаций к общему числу выполненных классификаций.

6. Результаты классификации для всех методов

На всех отрывках и mfw наилучшие результаты показали расстояние Лаббе (на абсолютных значениях частотности), косинусное сходство на стандартизированных

[9] В вопросе о подлинности речей Демосфена мы опирались на [Harris, 2013. P. 401–402], а также [Trevett, 2018]. Гиппократа и Лисия мы были вынуждены исключить из-за спорного статуса большинства сочинений корпуса. Так, [Dover, 1968. P. 193] признает безусловно подлинной лишь XII речь Лисия; критика у [Winter, 1973], но вопрос остается открытым.

[10] О способах обработки текстов в целом см.: [Stamatatos *et al.*, 2001; Stamatatos, 2009].

значениях, а также расстояние Танимото (относительная частотность без стандартизации). Чуть хуже показали себя Джеффрис и манхэттенское расстояние. Методы, давшие менее 0.6 процентов точности, исключены из дальнейшего рассмотрения.

Табл. 2. Средние показатели точности для всех методов

2.1. Словоформы

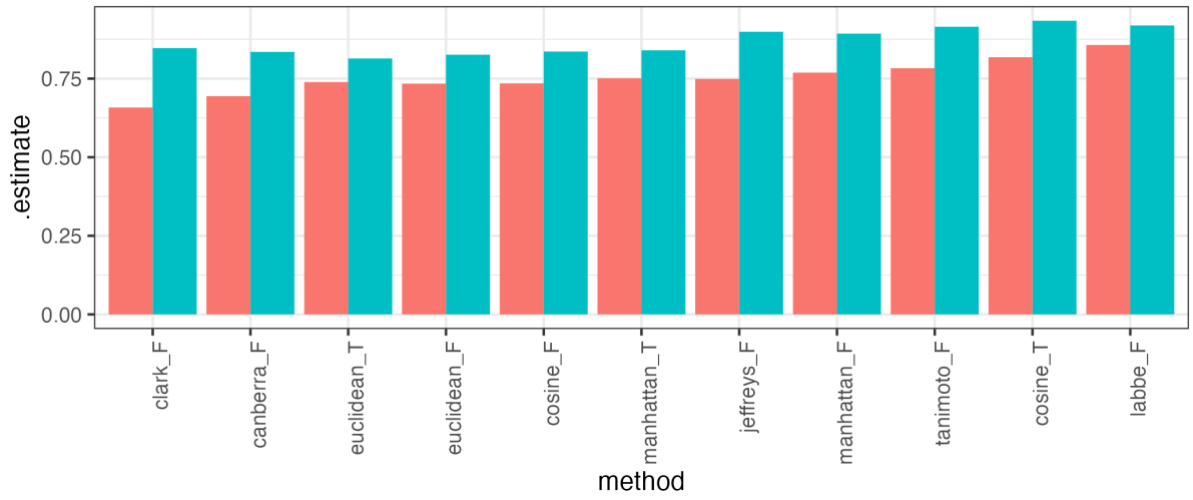
method	scale	.estimate
cosine	TRUE	0.934
labbe	FALSE	0.919
tanimoto	FALSE	0.915
jeffreys	FALSE	0.899
manhattan	FALSE	0.893
clark	FALSE	0.847
manhattan	TRUE	0.84
cosine	FALSE	0.836
canberra	FALSE	0.835
euclidean	FALSE	0.826
euclidean	TRUE	0.814
clark	TRUE	0.228
tanimoto	TRUE	0.208
canberra	TRUE	0.045

2.2. Энграмы

method	scale	.estimate
labbe	FALSE	0.857
cosine	TRUE	0.818
tanimoto	FALSE	0.783
manhattan	FALSE	0.769
manhattan	TRUE	0.751
jeffreys	FALSE	0.749
euclidean	TRUE	0.739
cosine	FALSE	0.735
euclidean	FALSE	0.734
canberra	FALSE	0.694
clark	FALSE	0.658
clark	TRUE	0.152
tanimoto	TRUE	0.11
canberra	TRUE	0.042

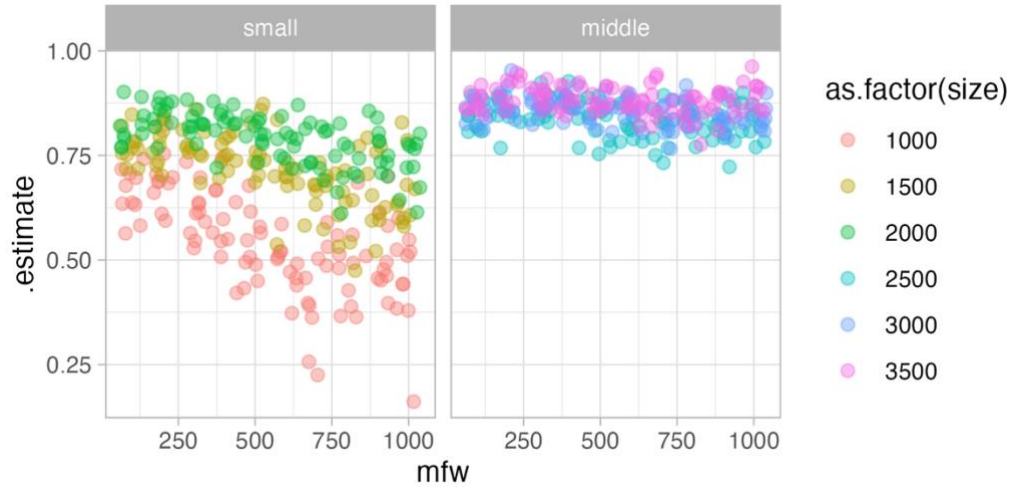
Классификация с использованием трехсимвольных энграм стабильно дает худший результат (Рис. 2), что может быть связано с особенностями древнегреческой диакритики.

Рис. 2. Точность классификации в зависимости от использования словоформ (бирюзовый) или энграм (розовый); T (True) указывает на стандартизацию данных



Как видно в Табл. 2, только косинусное сходство улучшает показатели со стандартизацией; таким образом, знаменитая Delta Берроуза [Орехов, 2020; Skorinkin, Orekhov, 2023] проигрывает не только альтернативным метрикам, но и манхэттенскому расстоянию без стандартизации, то есть фактически более примитивной версии самой себя.

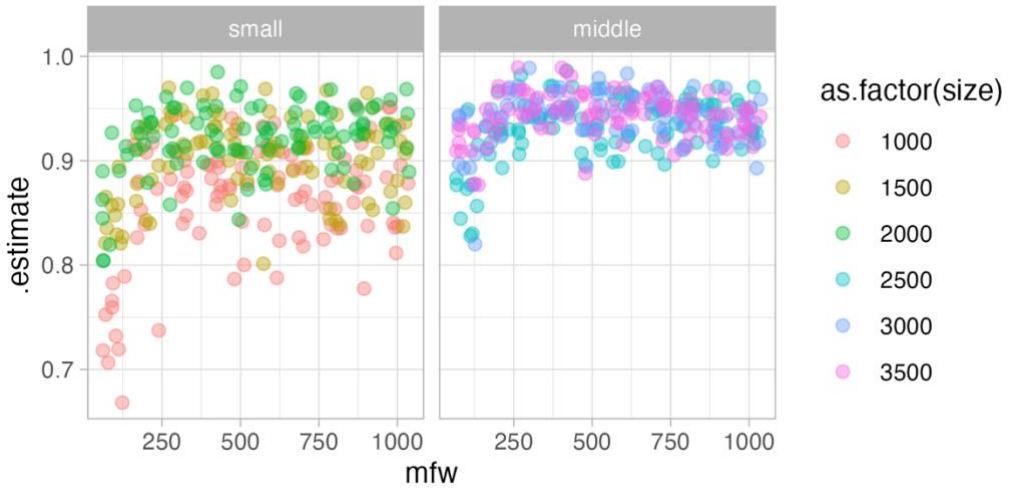
Рис. 3. Точность классификации для небольших и средних отрывков (Delta) в зависимости от числа mfw



Однако на Рис. 3 можно заметить, что показатели точности снижаются главным образом в тех случаях, когда количество mfw приближается к размеру отрывка, для чего Delta изначально не предназначена. На отрывках длиной 1000–2000 слов (слева) негативный эффект очевиден; для средних (справа) и тем более больших отрывков увеличение или уменьшение mfw уже не играет такой роли. Можно предположить, что

в случае с манхэттенским расстоянием без стандартизации увеличение числа mfw на небольших отрывках, хотя и ухудшает результат, но не так сильно — сказывается избыточное влияние гиперчастотных слов (согласно закону Ципфа). Для сравнения: косинусное сходство с применением стандартизации (далее COS_S) на небольших отрывках гораздо более устойчиво к изменениям mfw (Рис. 4), что отражается и на общей эффективности метрики.

Рис. 4. Точность классификации для небольших и средних отрывков (COS_S) в зависимости от числа mfw



7. Результаты классификации в зависимости от числа mfw

Рис. 5.1 и 5.2 подтверждают, что небольшие mfw имеют преимущество лишь при классификации с использованием расстояния Манхэттена на стандартизированных значениях (что соответствует подходу Берроуза). Однако это не распространяется на классификации с использованием трехсимвольных энграм. Для большинства методов точность либо возрастает вместе с числом mfw, либо не очень сильно от него зависит.

Рис. 5.1 Точность классификации в зависимости от числа mfw (словоформы)[11]

[11] Для удобства на графике линия, обозначенная как mfw 200, объединяет испытания на 100 и 200 mfw, и т.д.

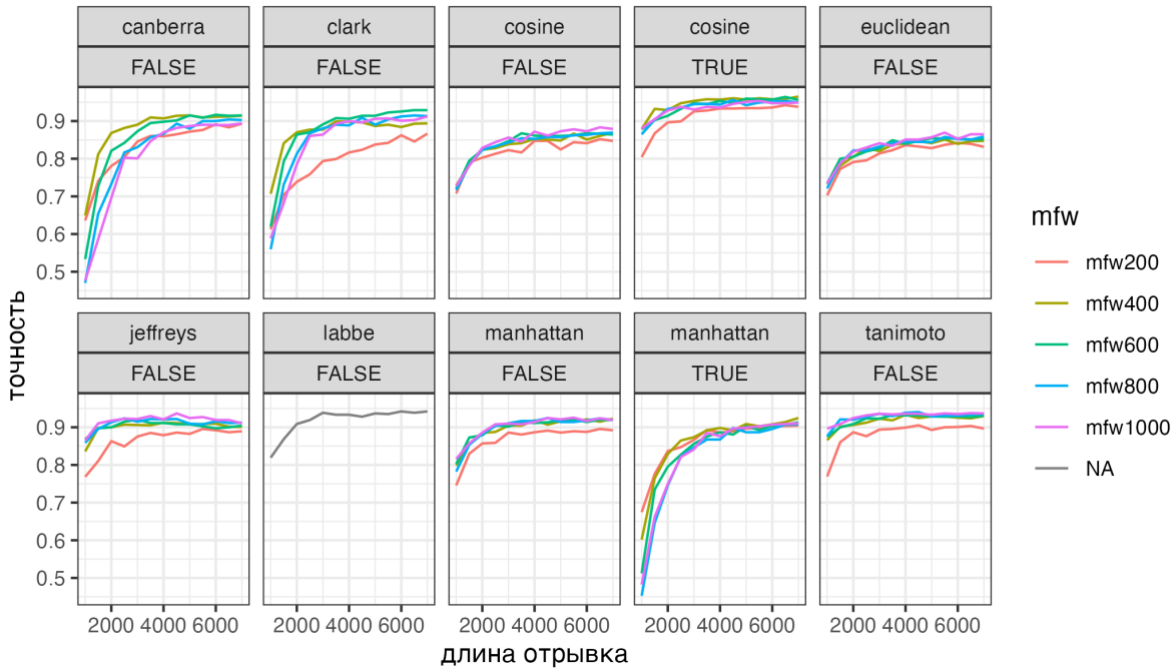
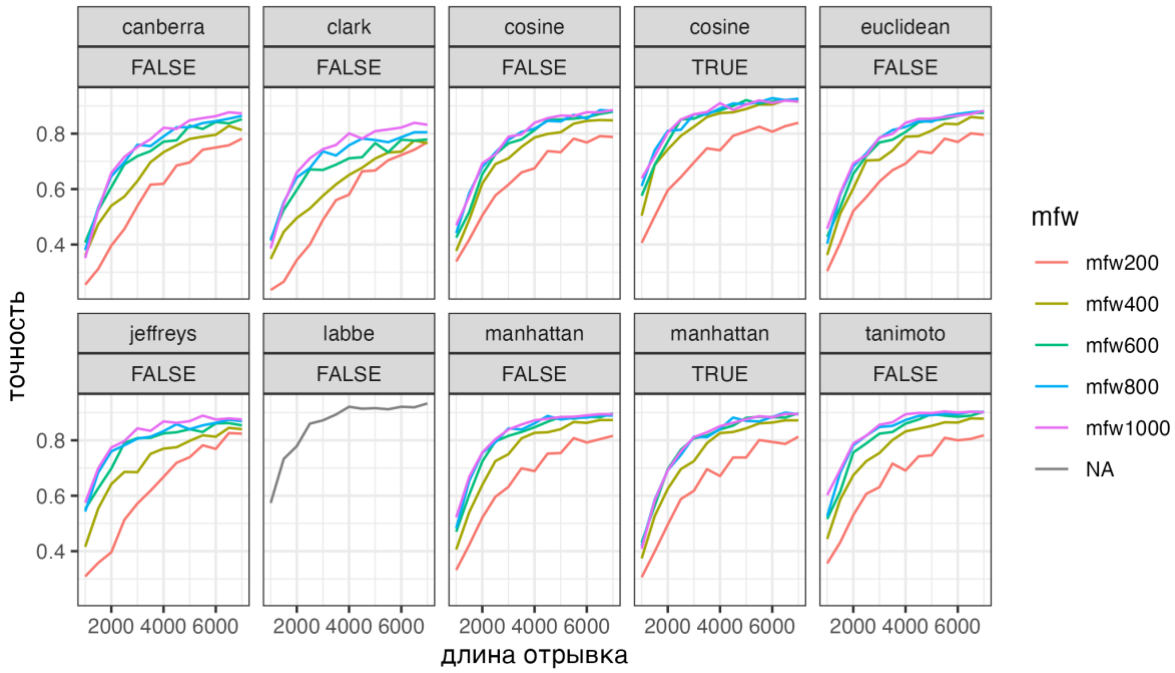


Рис. 5.2. Точность классификации в зависимости от числа mfw (трехсимвольные энграмы)



Наиболее стабильный результат достигается на словоформах при mfw > 200 и применении COS_S, расстояния Танимото (далее TAN), а также расхождения Джеффриса (далее JEFF). Для этих методов уже на отрывках в 2000-2500 токенов точность классификации стабилизируется на высоком уровне, что подтверждает

выводы Эдера [Eder, 2017]. В случае с расстоянием Лаббе (далее LAB) максимальная точность достигается начиная примерно с 3000 токенов.

При сопоставимой средней точности TAN, LAB и COS_S (см. выше Табл. 2) наилучшие результаты достигаются с применением именно последнего метода. В топе 50 лучших результатов (начало списка в Табл. 3) — только COS_S, причем на средних и больших отрывках, от 2500 токенов и выше.

Табл. 3. Наилучшие классификации с использованием словоформ (среднее значение для 10 итераций)

method	scale	mfw	size	mean
cosine	TRUE	500	6500	0.968
cosine	TRUE	300	7000	0.965
cosine	TRUE	400	7000	0.965
cosine	TRUE	500	7000	0.965
cosine	TRUE	600	5500	0.965
cosine	TRUE	300	4500	0.963
cosine	TRUE	400	6500	0.963
cosine	TRUE	300	5500	0.961
cosine	TRUE	300	6000	0.961
cosine	TRUE	400	3500	0.96
cosine	TRUE	400	4000	0.96
cosine	TRUE	400	5000	0.96
cosine	TRUE	400	5500	0.96
cosine	TRUE	500	5000	0.96
cosine	TRUE	600	5000	0.96
cosine	TRUE	600	6500	0.96
cosine	TRUE	200	6500	0.958
cosine	TRUE	400	4500	0.958
cosine	TRUE	500	6000	0.958

8. Методы на малых выборках

В исследовательской работе нередко приходится иметь дело с текстами меньшего размера, поэтому оптимальные параметры отобраны для отрывков в 1000, 1500 и 2000 слов. На отрывках 1000 точность нигде не превышает 0.9, причем TAN превосходит COS_S. На отрывках в 1500 и 2000 слов COS_S показывает весьма достойные результаты (Табл. 4).

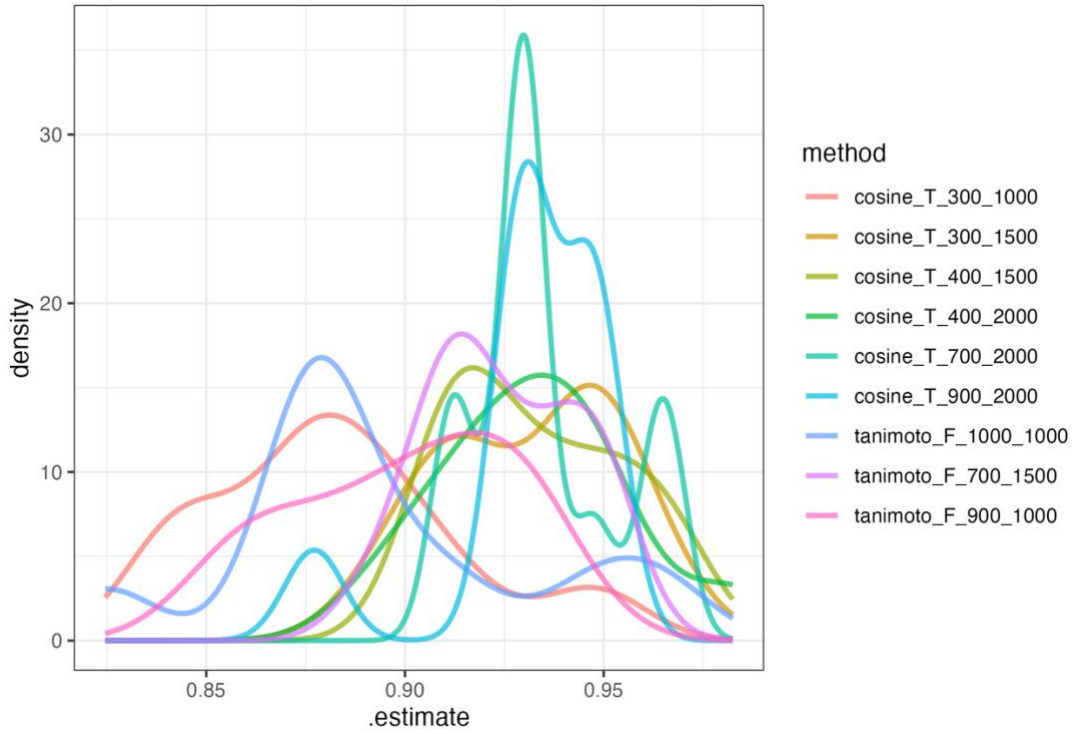
Табл. 4. Точность и стандартное отклонение точности на небольших отрывках (на основе 10 классификаций с повторными выборками)

method	scale	mfw	size	.estimate	.sd
tanimoto	FALSE	900	1000	0.9	0.027
tanimoto	FALSE	1000	1000	0.893	0.04
cosine	TRUE	300	1000	0.882	0.032
cosine	TRUE	400	1500	0.933	0.022
cosine	TRUE	300	1500	0.932	0.023
tanimoto	FALSE	700	1500	0.925	0.019

cosine	TRUE	700	2000	0.935	0.019
cosine	TRUE	400	2000	0.933	0.025
cosine	TRUE	900	2000	0.932	0.021

Наименьший разброс значений при высокой точности достигается, вполне ожидаемо, на отрывках 2000 слов. Это хорошо видно на Рис. 6, где отчетливо выделяются две «косинусные» вершины.

Рис. 6. Кривые плотности для лучших методов на малых выборках

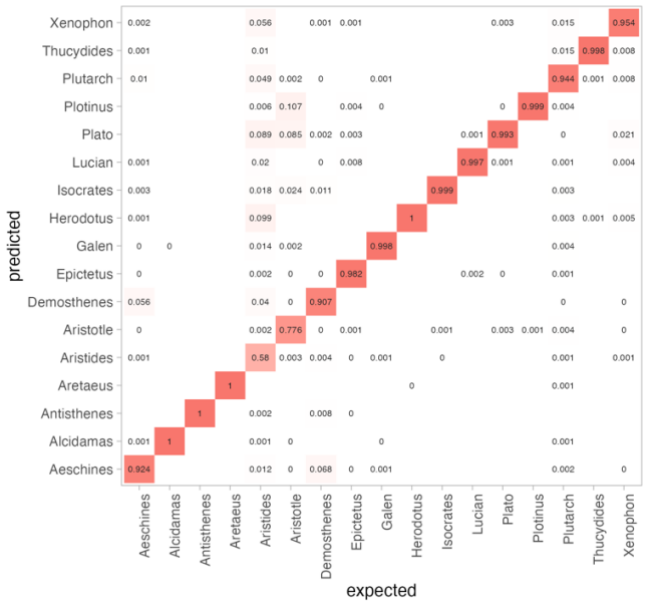


9. Матрицы ошибок

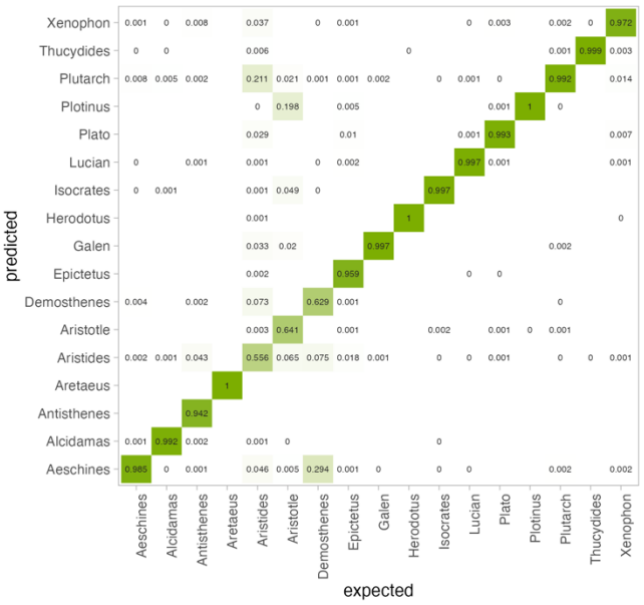
Поскольку ни в одном случае классификация не достигает стопроцентной точности, интересно узнать, на каких текстах чаще всего ошибается классификатор. Мы визуализировали матрицы ошибок для четырех методов: COS_S, TAN, JEFF, а также для Delta (Рис. 7).

Рис. 7. Матрицы ошибок

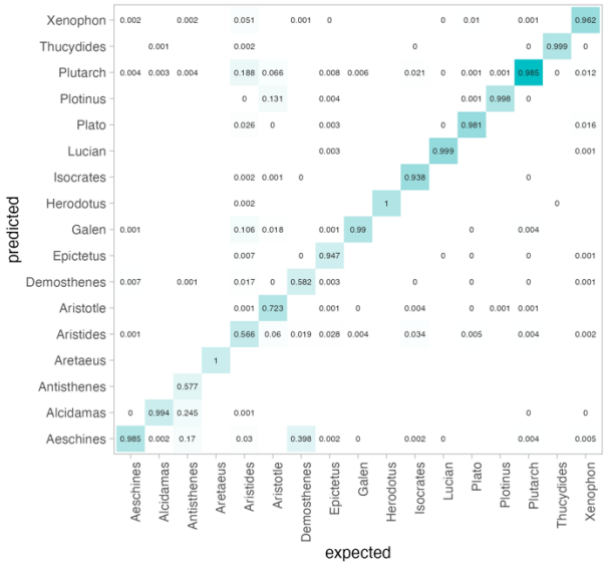
7.1. COS_S



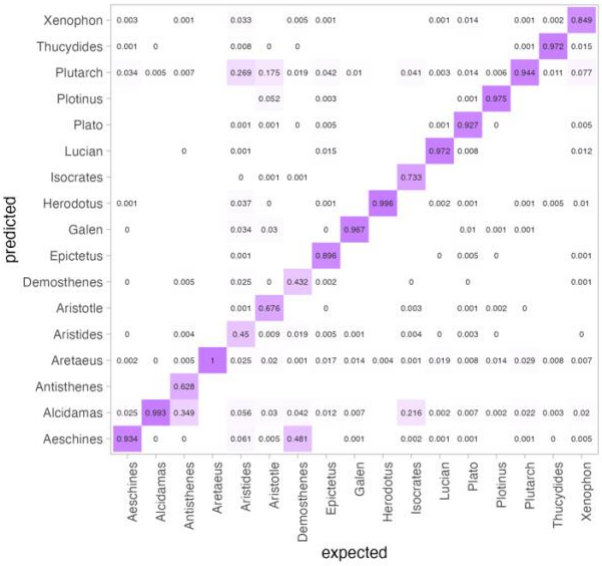
7.2. TAN



7.3. JEFF



7.4. Delta (MANH_S)



Для представленных методов наиболее «непредсказуемые» авторы совпадают — это Аристотель, Демосфен и Аристид; Delta и Джеффрис, кроме того, плохо справились с небольшими текстами Антисфена, отдав их отчасти Алкидаманту[12]. Delta также путает Исократа с Алкидамантом (что для обоих было бы очень обидно), Ксенофонта — с Фукидидом и Плутархом, а Эпикета — с Плутархом.

[12] LAB справляется с Антисфеном, но тоже спотыкается на Аристотеле, Демофене и Аристиде (точность примерно 0.7–0.75 для всех троих.

Из сочинений Плутарха в нашем корпусе были как моральные, так и исторические сочинения, и можно предположить, что смешение происходит между текстами близкой жанровой природы, что в целом характерно для Delta [Алиева, 2022]. Эти ошибки сохраняются, даже если удалить выборки размером 1000 токенов (см. выше раздел 6).

Ошибки, связанные с атрибуцией текстов Аристида, известного мастера стилизаций, и с путаницей между Демосфеном и Эсхином, более или менее понятны. Но судить о том, почему Аристотель оказывается похож на Плотина в каждом десятом случае (COS_S) или даже в каждом пятом (TAN), сложнее. С одной стороны, известно, что сочинения Аристотеля и его комментатора Александра Афродисийского читались в кружке Плотина (об этом сообщает Порфирий в *Vita Plotini* 14.13), и это должно было сказаться на философской терминологии «Эннеад». С другой стороны, отдельных общих терминов недостаточно для того, чтобы повлиять на классификатор, который пользуется сотнями предикторов[13]. В целом, вопрос как о стиле самого Аристотеля, так и о его отражении в позднеантичной литературе требует дополнительного изучения.

10. Тестирование

Наш небольшой эксперимент показал, что при классификации древнегреческих текстов наилучшие результаты достигаются при использовании косинусного сходства на стандартизированных значениях частотности, расстояния Танимото на относительных значениях частотности и расстояния Лаббе на абсолютных значениях частотности. Количество признаков для первых двух методов должно быть больше двухсот. Использование трехсимвольных энграм вместо словоформ ухудшает результат. Лаббе показывает высокую эффективность на отрывках от 3000 токенов, а на самых коротких отрывках лучше работает Танимото.

Можно ли говорить о статистически значимой разнице в точности предсказаний между этими тремя методами? Для сравнения трех векторов, хранящих данные о точности классификации на отрывках разной длины, мы использовали парный непараметрический тест Уилкоксона [Мастицкий, Шитиков, 2015. С. 157]. Для этого были удалены классификации с использованием mfw 100 и 200, показавшие наихудший результат, а результаты сохранены в виде трех векторов, хранящих

[13] Интерпретировать этот результат тем сложнее, что стиль Аристотеля (в отличие от стиля Платона) изучен очень плохо: помимо двух работ Ойкена [Eucken, 1866; Eucken 1868] можно упомянуть лишь исследования Кенни [Kenny, 1978. P. 70–160], но они посвящены почти исключительно «Этикам». Кроме того, как следует из диссертации Ойкена [Eucken, 1866. P. 15, 31 *et passim*], рукописные и издательские разночтения нередко касаются именно тех служебных слов, которые особенно важны как показатели авторского стиля (что, впрочем, верно для большинства античных авторов).

сведения о средней точности классификации по итогам десяти итераций на всех mfw (Табл. 5).

Табл. 5. Точность классификации (все mfw) для отрывков разной длины

COS_S

size	.estimate
1000	0.875
1500	0.911
2000	0.926
2500	0.938
3000	0.944
3500	0.946
4000	0.948
4500	0.953
5000	0.952
5500	0.955
6000	0.954
6500	0.956
7000	0.955

LAB

size	.estimate
1000	0.819
1500	0.868
2000	0.909
2500	0.919
3000	0.939
3500	0.933
4000	0.933
4500	0.928
5000	0.937
5500	0.935
6000	0.942
6500	0.939
7000	0.942

TAN

size	.estimate
1000	0.878
1500	0.908
2000	0.915
2500	0.924
3000	0.929
3500	0.929
4000	0.935
4500	0.933
5000	0.93
5500	0.931
6000	0.93
6500	0.931
7000	0.933

Тесты показывают отсутствие статистически значимой разницы между LAB и TAN, но на уровне значимости 0.05 отвергнута гипотеза об отсутствии различия между COS_S и LAB, а также TAN и COS. Таким образом, косинусное сходство оказывается наиболее эффективным методом на данном корпусе.

11. Применение к новым данным

Используя все три метода, мы попробовали классифицировать послание «К Демонику». Еще в древности высказывались сомнения в авторстве Исократа; сегодня принято считать, что текст создан одним из учеников Исократа в ответ на «Протрептик» Аристотеля[14]. Так, Фридрих Бласс полагал, что стиль «Демоника» недостоин (*nicht würdig*) Исократа [Blass, 1874. S. 257–258]. При этом Бласс учитывал то, что сам он называл *mikroskopische Eigenheiten*, такие как количество зияний или частотность предлога σύν[15], однако в случае с «Демоником» руководствовался,

[14] См.: [Mathieu, Brémond, 1928. P. 110–119] и [Düring, 1961. P. 23, 226]. [Wendland, 1905. S. 81-101] был первым, кто отметил параллели с Аристотелем; он также считает «Демоника» творением подражателя.
[15] [Blass, 1892. S. 295 = Бласс 2016. С. 193]: «Самыми убедительными являются мелкие признаки, те свойственные каждому писателю микроскопические особенности, которых подражатель или не мог

кажется, больше исследовательским чутьем, чем статистикой. Даг Хатчинсон, автор недавней реконструкции «Протрептика», настаивает, что убедительных доводов в пользу отвержения нет[16]. Используя три метода (COS_S size 6500, mfw 500; TAN size 6000 mfw 1000; LAB size 6000) и десять повторных выборок, мы сравнили это послание со всеми текстами в нашем корпусе. В результате все три метода отдали текст Исократу. Эти результаты следует интерпретировать осторожно: в нашем корпусе может не быть подлинного автора текста, и, если подражатель достаточно искусен, то атрибуция Исократу не удивительна.

Еще один пример — трактат «О воспитании детей» из корпуса Плутарха, который, согласно авторитетному суждению Виттенбаха, не может быть творением херонейского философа [Wyttenbach 1820. P. 1–30]. В этом случае наши метрики скорее подтверждают сомнения Виттенбаха: COS_S колеблется между Исократом и Плутархом (3/7); TAN — между Аристидом, Аристотелем и Плутархом (1/1/8); LAB — между Аристотелем и Аристидом (9/1). Это ничего не говорит нам о личности автора (подражателя скорее всего его нет в нашей коллекции), но в достаточной мере свидетельствует о том, что текст не очень похож на Плутарха.

И здесь уместно снова вернуться к Фридлендеру и его словам о «Пире», процитированным в начале этой статьи. Семь речей «Пира» написаны каждая в своей в узнаваемой стилистике, что должно создавать некоторые трудности при автоматической классификации текста[17]. Речь Агафона — блестящая имитация Горгия [Dover, 1968. P. 90–91], речь Федра, большого поклонника Лисия, похожа на речь самого Лисия из «Федра» [Tarrant, 2016. P. 86], а голос Алкивиада у Платона звучит во многом так же, как у Фукидида [Sansone, 2018]. Тем не менее, все три классификатора отдают «Пир» Платону: COS_S в 8 случаях из 10 (еще 2 — автору подложного «Феага», которого мы тоже добавили в подборку), LAB и TAN в 10 случаях из 10. Но справедливости ради надо сказать, что и «Феаг» оказывается ближе всего к Платону в

заметить, или подражание которым представилось бы слишком трудным. Таковы особенности относительно допущения зияния и соблюдения ритма, употребления известных частиц, избегания других, и т.п.». О вкладе Фридриха Бласса в количественные исследования стиля Платона см.: [Brandwood, 1990. P. 9–10].

[16] Письмо автору от 4.12.2017. Реконструкция доступна по ссылке: http://www.protrepticus.info/

[17] На подобные трудности указывалось и в связи с речами «Федра»; см., напр.: [Robinson, 1992. P. 381]. Вообще способность автора сознательно контролировать свой «почерк» — одно из главных возражений противников платоновской стилометрии в ее традиционном виде, то есть направленной на установление сравнительной хронологии диалогов; см., напр.: [Howland, 1991. P. 203] и [Waterfield, 1980. P. 275]. В то же время сознательная «смена регистра» все чаще становится предметом специальных количественных исследований; см., напр.: [Tarrant, Benitez, Roberts, 2011].

целом или к «Пиру» в частности (лишь в двух случаях TAN отдает текст Ксенофонту). С точки зрения статистики, это вполне удачная имитация, хотя *opinio communis* утверждает, что текст Платону не принадлежит, даже если и написан еще при жизни Платона [Thesleff, 2009. P. 364].

Таким образом, оптимистичные показатели точности не должны нас вводить в заблуждение, причем это касается не только описанных в этой статье методов, но и других, вероятно более совершенных. Авторитет традиции действителен до тех пор, пока не опровергнут, в то время как *подтвердить* его статистически не представляется возможным.

Литература

Алиева О.В. Delta Берроуза для древнегреческих авторов: опыт применения // Schole. Философское антиковедение и классическая традиция. 2022. Т. 16. № 2. С. 693–705.

Деза Е., Деза М.М. Энциклопедический словарь расстояний. М.: Наука, 2008.

Мастицкий С.Э., Шитиков В.К. Статистический анализ и визуализация данных с помощью R. М.: ДМК, 2015.

Орехов Б.В. «Илиада» Е.И. Кострова и «Илиада» А.И. Любжина: стилеметрический аспект // Аристей. Т. 21. 2020. С. 282–296.

Пешков И.В. Зарождение категории авторства в золотом веке английской литературы: дисс. … докт. филол. наук: 10.01.08. М.: РГГУ, 2016.

Фоминых С.В. Автоматический независимый от языка анализ авторства патристических текстов на основании статистики частот переходов // Исторический журнал: научные исследования. Т. 5. 2017. С. 70–79.

Фуко М. Что такое автор? // Эстетика и теория искусства XX в.: Хрестоматия / Н.А. Хренов, А.С. Мигунов (сост.). М.: Прогресс-традиция, 2008.

Хачумов М.В. Расстояния, метрики и кластерный анализ // Искусственный интеллект и принятие решений. Т. 9. № 1. 2012. С. 81–89.

Шумилин М.В. Ошибка против варианта: «новая филология», латинистика и «плохой язык» // Vox medii aevi. Т. 1–2. 2020. С. 28–75. <http://voxmediiaevi.com/volumens/2020-1-2/2020-1-2-shumilin/>

Ahmed H. Distance-Based Authorship Verification Across Modern Standard Arabic Genres // Proceedings of the Third Workshop on Arabic Corpus Linguistics. 2019. P. 89–96. <https://aclanthology.org/W19-5611.pdfn>.

Argamon S. Interpreting Burrows' Delta: Geometric and Probabilistic Foundations // Literary and Linguistic Computing. Vol. 23. No. 2. 2008. P. 131–147.

Blass F.W. Attische Beredsamkeit. 2. Abtheilung: Isokrates und Isaios. Leipzig: B.G. Teubner, 1874.

Blass F.W. Hermeneutik und Kritik // Handbuch der klassischen Altertums-Wissenschaft in systematischer Darstellung mit besonderer Rücksicht auf Geschichte und Methodik der einzelnen Disziplinen / I. von Müller (hrsg.). Bd. 1: Einleitende und Hilfs-Disziplinen. 2. Aufl. München: C.H. Beck, 1892. Бласс Ф.В. Герменевтика и критика: искусство понимания произведений классической древности и их литературная оценка / Л.Ф. Воеводский (пер.). 2-е изд. М.: Ленанд, 2016.

Boeckh A. Encyklopädie und Methodologie der philologischen Wissenschaften / Ernsts Bratuscheck (hrsg). 2. Aufl. Wiesbaden: Springer Fachmedien, 1886.

Brandwood L. The Chronology of Plato's Dialogues. Cambridge: Cambridge University Press, 1990.

Burrows J.F. 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship // Literary and linguistic computing. Vol. 17. No. 3. 2002. P. 267–287.

Chambers M. The *Athenaion Politeia* after a Century // Transitions to Empire. Essays in Greco-Roman History, 360– 140 B.C., in honor of E. Badian / R. Wallace and E. Harris (eds.). Norman and London: University of Oklahoma Press, 1996. P. 211–225.

Cortelazzo M.A., Nadalutti P., Tuzzi A. Improving Labbé's Intertextual Distance: Testing a Revised Version on a Large Corpus of Italian Literature // Journal of Quantitative Linguistics. Vol. 20. No. 2. 2013. P. 125–152.

Dover K.J. Lysias and the *Corpus Lysiacum*. Berkeley and Los Angeles: University of California Press, 1968.

Eder M. Style-Markers in Authorship Attribution A Cross-Language Study of the Authorial Fingerprint // 2011. No 6. P. 99–114.

Eder M. Taking Stylometry to the Limits: Benchmark Study on 5,281 Texts from "Patrologia Latina" // Digital Humanities 2015: Book of Abstracts, University of Western Sydney. : Alliance of Digital Humanities Organizations (ADHO), 2015.

Eder M. Short Samples in Authorship Attribution: A New Approach // Digital Humanities 2017. Montreal, 2017. https://dh2017.adho.org/abstracts/341/341.pdf

Eucken R. De Aristotelis dicendi ratione: Pars prima. Observations de particularum usu. Dissertatio inauguralis. Gottingae: Typis expressit officina Hoferiana, 1866.

Eucken R. Ueber den Sprachgebrauch des Aristoteles: Beobachtungen ueber die Praepositionen. Berlin: Weidmansche Buchhandlung, 1868.

Evert S., Proisl Th., Jannidis F., Reger I., Pielström S., Schöch Ch., Vitt Th. Understanding and explaining Delta measures for authorship attribution // Digital Scholarship in the Humanities. Vol. 32. Supp. 2. 2017. P. ii4-ii16.

Forcignanò F., Martinelli Tempesta S. Comparing Corpora, Rethinking Authenticity: Why Are Platonic Letters "Platonic"? // The Making of the Platonic Corpus / Alieva O., Nails D., Tarrant H. (eds.). Paderborn: Brill, 2023. P. 203–221.

Harris E.M. The Rule of Law in Democratic Athens. Oxford: Oxford University Press, 2013.

Holmes D.I. Authorship Attribution // Computers and the Humanities. Vol. 28. No. 2. 1994. P. 87–106.

Hoover D.L. Testing Burrows's Delta // Literary and Linguistic Computing. Vol. 19. No. 4. 2004. P. 453– 475.

Howland J. Re-Reading Plato: The Problem of Platonic Chronology // Phoenix. Vol. 45. No. 3. 1991. P. 189–214.

Jannidis F., Lauer G. Burrows's Delta and Its Use in German Literary History // Distant Readings. Topologies of German Culture in the Long Nineteenth Century Studies in German Literature Linguistics and Culture / M. Erlin, L. Tatlock (eds.). Rochester: Camden House, 2014. P. 29–54.

Juola P. Cross-Entropy and Linguistic Typology // New Methods in Language Processing and Computational Natural Language Learning / D.M.W. Powers (ed.). Sydney: ACL, 1998. P. 141–149. <https://aclanthology.org/W98-1217.pdf>

Juola P. How a Computer Program Helped Show J.K. Rowling write *A Cuckoo's Calling*: Author of the Harry Potter books Has a Distinct Linguistic Signature // Scientific American 20.08.2013 <https://www.scientificamerican.com/article/how-a-computer-program-helped-show-jk-rowling-write-a-cuckoos-calling/>

Juola P. What Can We Do with Small Corpora? Document Categorization via Cross-entropy // Proceedings of an Interdisciplinary Workshop on Similarity and Categorization. Edinburgh: University of Edinburgh, 1997.

Juola P., Baayen H. A Controlled-corpus Experiment in Authorship Identification by Cross-Entropy // Literary and Linguistic Computing. Vol. 20 Suppl. 2005. P. 59–67.

Kenny A. The Aristotelian Ethics: A Study of the Relationship between the *Eudemian* and *Nicomachean Ethics* of Aristotle. Oxford: Clarendon Press: 1978.

———. 1982. The Computation of Style: An Introduction to Statistics for Students of Literature and Humanities. Oxford, Pergamon Press.

Kestemont M., Stover J., Koppel M., Karsdorp F., Daelemans W. Authenticating the writings of Julius Caesar // Expert Systems with Applications. Vol. 63. 2016. P. 86–96.

Kocher M., Savoy J. Evaluation of Text Representation Schemes and Distance Measures for Authorship Linking // Digital Scholarship in the Humanities. Vol. 34. No. 1. 2019. P. 189–207.

Koentges Th. The Un-Platonic *Menexenus*: A Stylometric Analysis with More Data // Greek, Roman, and Byzantine Studies. Vol. 60. 2020. P. 211–241.

Koppel M., Winter Y. Determining If Two Documents are Written by the Same Author // Journal of the Association for Information Science and Technology. Vol. 65. No. 1. 2014. P. 178–187.

Labbé D. Experiments on Authorship Attribution by Intertextual Distance in English // Journal of Quantitative Linguistics. Vol. 14. No. 1. 2007. P. 33–80.

Labbé C., Labbé D. A Tool for Literary Studies: Intertextual Distance and Tree Classification // Literary and Linguistic Computing. Vol. 21. No. 3. 2006. P. 311–326.

Ledger G. Recounting Plato: A Computer Analysis of Plato's Style. Oxford: Oxford University Press, 1989.

Love H. Attributing Authorship: An Introduction. Cambridge: Cambridge University Press, 2002.

Mathieu G., Brémond E. eds. 1928. Isocrate: Discours. T. 1. Paris: Les belles lettres.

Robinson T.M. Plato and the Computer // Ancient Philosophy. Vol. 12. 1992. P. 375–382.

Rudman J. Unediting, De-Editing, and Editing in Nontraditional Authorship Attribution Studies: With an Emphasis on the Canon of Daniel Defoe // The Papers of the Bibliographical Society of America. Vol. 99. No. 1. 2005. P. 5–36.

Rybicki J., Eder M. Deeper Delta across Genres and Languages: Do We Really Need the Most Frequent Words? // Literary and Linguistic Computing. Vol. 26. No 3. 2011. P. 315–321.

Sansone D. Stylistic Characterization in Plato: Nicias, Alcibiades, and Laches // Greek, Roman, and Byzantine Studies. Vol. 58. 2018. P. 156–176.

Savoy J. Machine Learning Methods for Stylometry: Authorship Attribution and Author Profiling. Cham: Springer, 2020.

Shannon C.E. A Mathematical Theory of Communication // The Bell System Technical Journal. Vol. 27. No. 3. 1948. P. 379–423.

Skorinkin D., Orekhov B. Hacking stylometry with multiple voices: Imaginary writers can override authorial signal in Delta // Digital Scholarship in the Humanities. 2023 <https://doi.org/10.1093/llc/fqad012>

Smith P.W.H., Aldridge W. Improving Authorship Attribution: Optimizing Burrows' Delta Method // Journal of Quantitative Linguistics. Vol. 18. No. 1. 2011. P. 63–88.

Stamatatos E. A Survey of Modern Authorship Attribution Methods // Journal of the American Society for Information Science and Technology. Vol. 60. No. 3. 2009. P. 538–556.

Stamatatos E., Kokkinakis G., Fakotakis N. Automatic Text Categorization in Terms of Genre and Author // Computational linguistics. Vol. 26. No. 4. 2000. P. 471–495.

Stanikūnas D., Mandravickaitė J., Krilavičius T. Comparison of distance and similarity measures for stylometric analysis of Lithuanian texts // ICYRIME 2017: Proceedings of the Symposium for Young Researchers in Informatics, Mathematics and Engineering. Aachen: CEUR-WS, 2017. P. 1–7 <https://ceur-ws.org/Vol-1852/p01.pdf>

Tarán L. Academica: Plato, Philip of Opus, and the Pseudo-Platonic *Epinomis*. Philadelphia: American Philosophical Society, 1975.

Tarrant H. Stylistic Difference in the Speeches of the *Symposium* // Plato in Symposium: Selected Papers from the Tenth Symposium Platonicum / M. Tulli and M. Erler (eds.). Sankt Augustin: Academia Verlag, 2016. P. 84–90.

Tarrant H., Benitez E.E., Roberts T. The Mythical Voice in the *Timaeus-Critias*: Stylometric Indicators // Ancient Philosophy. Vol. 31. 2011. P. 95–120.

Thesleff H. Platonic Patterns: A Collection of Studies by Holger Thesleff. Las Vegas, Zurich, Athens: Parmenides Publishing, 2009.

Thesleff H. Afterthoughts on "School Accumulation" in Plato's Academy // The Making of the Platonic Corpus / Alieva O., Nails D., Tarrant H. (eds.). Paderborn: Brill, 2023. P. 1–14.

Trevett J. 2018. Authenticity, Composition, Publication // The Oxford Handbook of Demosthenes / G. Martin (ed.). Oxford. P. 419–430.

Waterfield R. A. H. The Place of the *Philebus* in Plato's Dialogues // Phronesis. Vol. 25, No. 3. 1980. P. 270-305.

Wendland P. Anaximenes von Lampsakos: Studien zur ältesten Geschichte der Rhetorik // Festschrift für die XLVIII. Versammlung deutscher Philologen und Schulmänner in Hamburg, von Paul Wendland. Berlin, 1905.

Winter T.N. On the Corpus of Lysias // The Classical Journal. Vol. 69. No. 1. 1973. P. 34–40.

Wyttenbach D. Animadversiones in Plutarchi *Opera moralia*. Lipsiae, 1820.

Zhao Y., Zobel J. Entropy-Based Authorship Search in Large Document Collections // Lecture Notes in Computer Science. Vol. 4425. 2007. P. 381–392.

Zhao Y., Zobel J., Vines Ph. Using Relative Entropy for Authorship Attribution // Lecture Notes in Computer Science. Vol. 4182. 2006. P. 92–105.