Меры расстояния для определения авторства древнегреческих текстов

Аннотация

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 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]. По умолчанию традиция достоверна, а бремя доказательства лежит на том, кто ее отвергает: *quivis praesumitur genuinus liber, donec demonstretur contrarium* [Boeckh, 1886. S. 240][1].

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

[1] «Всякая книга признается подлинной, пока не доказано обратное» (лат.).

вмешательств понятно [Rudman, 2005], но применительно к древним текстам едва ли выполнимо[2]. 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 значения непрерывной величины, между которыми есть еще «почти автор» и «не вполне автор» [Thesleff, 2009. P. 351]. 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

[2] О проблеме «неустойчивости» античных и средневековых текстов и подходах к их реконструкции см.: [Шумилин, 2020. С. 33 с прим. 11 *et passim*].

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: арабском [Ahmed, 2019], литовском [Stanikūnas *et al*., 2017], латинском [Kestemont *et al*., 2016] и многих других. 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.

Меры расстояния

Для сравнения 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[3] [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.

[3] Также называемое подобность Орчини, угловая подобность или нормированное скалярное произведение [Деза, Деза, 2008. С. 264].

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[4]. 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	$D_{Euclidean}$	$\sqrt{\sum_{i=1}^{n} \lvert P_i - Q_i \rvert^2}$
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}	$\sqrt{\sum_{i=1}^{n} \left(\frac{\lvert P_i - Q_i \rvert}{P_i + Q_i} \right)^2}$

[4] 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].

7	$D_{Jeffreys}$	$\sum_{i=1}^{n}(P_i - Q_i)ln\frac{P_i}{Q_i}$
8	$D_{Labbé}$	$\frac{\sum_{i=1}^{n}\lvert P_i - Q_i\rvert}{2N_P}$

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 трехсложных энграм;
4) сделать выводы о том, на каких текстах классификатор чаще ошибается;

Все вычисления и визуализации выполнены в R, при этом токенизация (деления 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[5].

Корпус

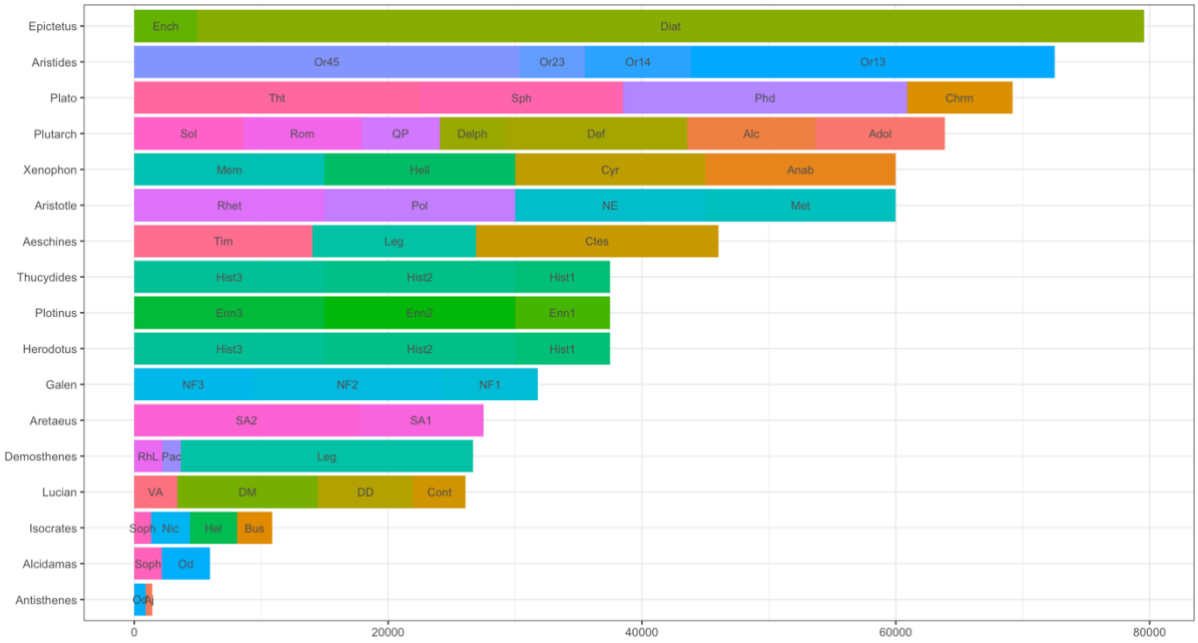
В корпус, подготовленный для этого эксперимента, вошли сочинения древнегреческих прозаиков (историков, врачей, философов и ораторов) из библиотеки Perseus[6]. Объем корпуса —694 тыс. слов. Корпус является несбалансированным в двух отношениях:

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

[6] 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).

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 полутора тысяч 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 «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»)[7]. Итого 57 текстов (включая выборки) и 17 авторов. Тексты не подвергались лемматизации, стеммингу, частеречной или синтаксической разметке[8]. Для анализа они были только разделены на токены: словоформы и трехбуквенные энграмы (с сохранением диакритики).

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 проводено 10 итераций, при этом использовались повторные выборки (с замещением из-за наличия в корпусе очень коротких текстов). Таким образом, для оценки *каждого* метода выполнено $57 \cdot 10 \cdot 13 \cdot 10 = 74100$ 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.

Результаты классификации

На всех отрывках и mfw наилучшие результаты показали расстояние Лаббе (на абсолютных значениях частотности), косинусное сходство на стандартизированных значениях, а также расстояние Танимото (относительная частотность без стандартизации). При этом только косинусное сходство дает лучшие результаты со стандартизацией; методы, давшие менее 0.6 процентов точности, исключены из дальнейшего рассмотрения.

---

[7] В вопросе о подлинности речей Демосфена мы опирались на [Harris, 2013. P. 401–402]. 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.

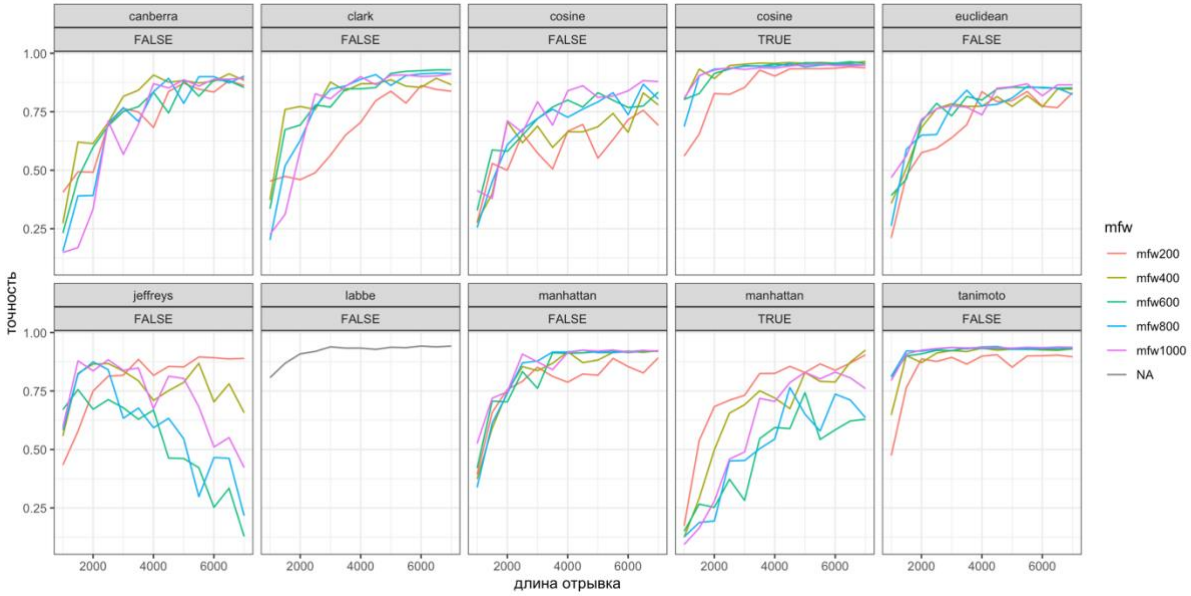
[8] О способах обработки текстов в целом см.: [Stamatatos *et al*., 2001; Stamatatos, 2009].

Табл. 2. Средние показатели точности для всех методов

метод	scale	точность
labbe	FALSE	0.918
cosine	TRUE	0.915
tanimoto	FALSE	0.901
manhattan	FALSE	0.821
clark	FALSE	0.765
euclidean	FALSE	0.732
canberra	FALSE	0.721
jeffreys	FALSE	0.68
cosine	FALSE	0.675
manhattan	TRUE	0.595
euclidean	TRUE	0.258
clark	TRUE	0.115
canberra	TRUE	0.053
tanimoto	TRUE	0.042

Как видно на Рис. 2, для большинства методов точность возрастает вместе с числом mfw (для удобства на графике линия, обозначенная как mfw 200, объединяет испытания на 100 и 200 mfw, и т.д.). Небольшие mfw имеют преимущество лишь при классификации с использованием расстояния Джеффриса и расстояния Манхеттена на стандартизированных значениях (что соответствует подходу Берроуза). Однако наиболее стабильный результат достигается при mfw > 200 и применении косинусного сходства (со стандартизацией, далее COS_S) или расстояния Танимото (далее TAN). Для этих методов уже на отрывках в 2000-2500 слов точность классификации стабилизируется на высоком уровне, что подтверждает выводы Эдера [Eder, 2017]. Похожая картина наблюдается при классификации с использованием дистанции Лаббе (далее LAB).

Рис. 2. Точность классификации в зависимости от длины отрывка и числа слов



При сопоставимой средней точности TAN, LAB и COS_S (см. Табл. 2) наилучшие результаты достигаются с применением именно последнего метода (Табл. 3).

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

method	scale	mfw	size	.estimate
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

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; а вот на отрывках в 1500 и 2000 слов COS_S показывает весьма неплохие результаты (Табл. 4).

Табл. 4. Точность на небольших отрывках

method	scale	mfw	size	.estimate
cosine	TRUE	700	2000	0.935
cosine	TRUE	400	1500	0.933
cosine	TRUE	400	2000	0.933
cosine	TRUE	300	1500	0.932
tanimoto	FALSE	800	1000	0.882
cosine	TRUE	400	1000	0.875

Стандартное отклонение точности для методов COS_S (mfw 700, size 2000), COS_S (mfw 400, size 1500) и TAN (mfw 800, 1000) составляет, соответственно, 0.018, 0.021 и 0.024 (на основе 10 классификаций с повторными выборками).

Поскольку ни в одном случае классификация не достигает стопроцентной точности, интересно узнать, на каких текстах чаще всего случается ошибка. Для трех наилучших методов (LAB, TAN, COS_S) наиболее «непредсказуемые» авторы совпадают —это

Аристотель, Демосфен и Аристид. Табл. 5 позволяет судить о том, с кем чаще всего путает этих авторов классификатор (COS_S, все mfw, все размеры).

Табл. 5. Ошибки классификации для трех авторов (выбраны max 3 «двойника»)

expected	predicted	n
Aristides	Aristides	2999
Aristides	Herodotus	511
Aristides	Plato	461
Aristides	Xenophon	287
Aristotle	Aristotle	3959
Aristotle	Plotinus	534
Aristotle	Plato	455
Aristotle	Isocrates	112
Demosthenes	Demosthenes	3466
Demosthenes	Aeschines	264
Demosthenes	Aristotle	73
Demosthenes	Isocrates	36

Привлекает внимание частая атрибуция отрывков из Аристида Геродоту, а также смешение между Аристотелем и, достаточно неожиданно, Исократом. Аристид, как многие авторы «второй софистики», — мастер стилизаций, что может объяснять подобные ошибки, а в случае Аристотеля трудность, как мы полагаем, связана с включением в корпус четырех различных по тематике его трактатов. Любопытно, что небольшие тексты Антисфена и Алкидаманта совершенно не затерялись в корпусе (точность 0.99–1 для LAB, TAN, COS_S) — своеобразное подтверждение их своеобычного стиля. В целом, любой из трех отмеченных методов дает результаты лучшие, чем Delta Берроуза.

Добавим, что при этом для всех трех методов точность ниже, если использовать трехбуквенные энграммы: LAB 0.801, COS_S 0.665, TAN 0.638. Точность классификации с применением остальных дистанций при таком подходе еще ниже.

Заключение

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 трехбуквенных энграм вместо словоформ ухудшает результат.

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 сведения о средней точности классификации по итогам десяти итераций на всех mfw (Табл. 6).

Табл. 6. Точность классификации (все mfw) для отрывков разной длины

COS_S

size	.estimate
1000	0.774
1500	0.892
2000	0.917
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	0.807
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	0.766
1500	0.908
2000	0.906
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

Тест показал 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 (значения средних см. выше Табл. 2). Таким образом, косинусное сходство оказывается наиболее эффективным методом при условии, что mfw > 200 (заметь разницу с Табл. 2).

Используя все три метода, мы попробовали классифицировать послание «К Демонику». Большинство исследователей считают, что оно принадлежит не самому

Исократу, а одному из его учеников, однако известный исследователь античности Даг Хатчинсон (Тринити-колледж, Торонто), считает, что убедительных доводов в пользу отвержения нет.[9] Используя три метода (COS_S size 6500, mfw 500; TAN size 6000 mfw 1000; LAB size 6000), мы сравнили это послание со всеми текстами в нашем корпусе. В результате 10 классификаций из 10 с использованием TAN и LAB отдали текст Геродоту; COS_S лишь в 5 случаях из 10 приписала текст Исократу. Эти результаты следует интерпретировать осторожно: в нашем корпусе может не быть подлинного автора текста (возможно, «ученик Исократа» больше вообще ничего не написал). Если бы текст был приписан Исократу, это не означало бы *подтверждения* авторства; как мы сказали в начале этой статьи, такое подтверждение вообще невозможно. Если подражатель достаточно искусен, то нет никаких сомнений, что из множества текстов подражание окажется ближе всего к своему образцу. Но вот атрибуция Геродоту в большинстве случаев — повод усомниться в том, что текст действительно написан Исократом.

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.

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.

[9] Письмо автору от 4.12.2017.

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. 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.

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

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.

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.

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.

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.

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.

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.

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>

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

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

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

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.