ТЕОРЕТИЧЕСКАЯ, ПРИКЛАДНАЯ И СРАВНИТЕЛЬНО-СОПОСТАВИТЕЛЬНАЯ ЛИНГВИСТИКА / THEORETICAL, APPLIED AND COMPARATIVE LINGUISTICS

DOI: https://doi.org/10.18454/RULB.2023.45.28

ЧИСЛИТЕЛЬНЫЕ В ТЕКСТАХ КАК ХАРАКТЕРНАЯ ОСОБЕННОСТЬ АВТОРСКОГО СТИЛЯ

Научная статья

Зенков А.В.[1, *], Ермаков Н.Е.[2]
[1] ORCID : 0000-0002-1233-9082;
[1, 2] Уральский федеральный университет, Екатеринбург, Российская Федерация

* Корреспондирующий автор (zenkow[at]mail.ru)

Аннотация

Рассмотрено применение нового количественного метода изучения авторского стиля литературных текстов. Метод основан на компьютерном анализе статистики числительных, встречающихся в текстах. Показано, что количественные и порядковые числительные, используемые автором в (художественном) тексте, индивидуальны для каждого автора; их совокупность является характерным признаком, различающим тексты разных авторов. Выполнен сопоставительный анализ литературных текстов И.А. Бунина и А.И. Куприна; для проверки методологии дополнительно исследованы произведения Ф.К. Сологуба, М.П. Арцыбашева. Результаты анализа подвергнуты иерархической кластеризации, правильно разделившей тексты в соответствии с авторством и жанром.

Ключевые слова: стилометрия, квантитативная лингвистика, атрибуция текстов, авторство текстов, числительные в тексте.

NUMERALS IN TEXTS AS A CHARACTERISTIC TRAIT OF AUTHOR'S STYLE

Research article

Zenkov A.V.[1, *], Ermakov N.E.[2]
[1] ORCID : 0000-0002-1233-9082;
[1, 2] Ural Federal University, Ekaterinburg, Russian Federation

* Corresponding author (zenkow[at]mail.ru)

Abstract

The application of a new quantitative method of studying the author's style of literary texts is examined. The method is based on computer analysis of the statistics of numerals occurring in texts. It is shown that quantitative and ordinal numerals used by the author in the (fiction) text are individual for each author; their aggregate is a characteristic trait that distinguishes the texts of different authors. A comparative analysis of literary texts by I.A. Bunin and A.I. Kuprin is carried out; to verify the methodology, the works of F.K. Sologub and M.P. Artsybashev are additionally studied. The results of the analysis were subjected to hierarchical clustering, which correctly divided the texts according to authorship and genre.

Keywords: stylometry, quantitative linguistics, text attribution, text authorship, numerals in text.

Введение

Задачи стилометрии, к которым относится количественное изучение авторских особенностей текстов (в т.ч. для их атрибуции) до настоящего времени не имеют вполне удовлетворительного решения [1], [2]: традиционно вычисляются частоты встречаемости в текстах знаменательных частей речи и служебных слов (предлоги, союзы), средние длины слов и предложений; в паре анализируемых текстов сравниваются самые часто встречающиеся слова [3] и даже буквосочетания (как ни странно, последний подход часто даёт неплохие результаты). К сожалению, разные методы часто приводят к несогласующимся выводам.

Хорошие результаты получены с помощью нейронных сетей [4], а вскоре, по-видимому, искусственный интеллект сможет успешно решать задачи стилометрии, но содержательная интерпретация результатов при этом затруднительна, поскольку метод, опирающийся на применение нейронных сетей, является «чёрным ящиком».

Нами разработан оригинальный подход к анализу авторских (литературных) текстов, основанный на учёте использования авторами числительных в их текстах [5], [6]. Такой подход имеет немалые преимущества. Среди знаменательных частей речи числительные по своей природе наиболее легко поддаются количественному учёту. Кроме того, среди всех особенностей текста, анализируемых в стилометрии, пожалуй, только встречаемость числительных практически не меняется при переводе текста на другой язык (за малозначимым вычетом числительных, входящих в идиомы – см. ниже). Это расширяет возможности текстологического анализа, позволяя, например, в случае необходимости привлекать тексты, имеющиеся на языке-посреднике.

Применительно к художественному тексту, содержание которого не является жёстко привязанным к реальным событиям, а порождено свободным воображением, естественно предположить, что употребление числительных связано с психологическими особенностями автора, неосознанно для него самого влияющими на результат творчества (в рамках выбранных жанра, сюжета и т.п.). Следовательно, манера использования числительных – это авторская особенность (fingerprint), позволяющая при определённых обстоятельствах решить проблему авторства текста.

Наше предположение подтвердилось; анализ произведений нескольких десятков авторов на русском, чешском, английском языках обнаружил ощутимые авторские особенности употребления числительных в текстах, влияние на них жанра, стиля, художественного направления [7], [8], [9], [10]. В частности, недавно нами решена важная проблема

чешского литературоведения, связанная с авторством некоторых текстов, приписываемых классику чешской литературы – Петру Безручу [10]. Таким образом, результаты анализа встречаемости числительных допускают содержательное филологическое истолкование.

В данной работе мы проанализируем основные литературные произведения И.А. Бунина (1870–1953) и А.И. Куприна (1870–1938) с точки зрения использования числительных. Литературная критика часто рассматривает эти имена совместно, а в художественной манере Нобелевского лауреата по литературе Бунина и его современника Куприна, которого Бунин высоко ценил, находят немало общего [11], [12], [13]. До сих пор сопоставление ограничивалось традиционными филологическими подходами.

В современной стилометрии укоренилось мнение, что даже при сопоставлении текстов двух авторов доказательную силу об их сходстве будет иметь лишь анализ, в котором изучаемые тексты «разбавлены» множеством посторонних текстов других авторов (т.н. impostors) [14]. Следуя этим идеям, мы ввели в рассмотрение произведения двух современников Бунина и Куприна – Ф.К. Сологуба (1863–1927) и М.П. Арцыбашева (1878–1927). Выбор диктовался частичным совпадением художественных направлений всех четырёх литераторов [11], [12], [13] и личными предпочтениями автора настоящего исследования.

Метод и объекты исследования

Нами разработана компьютерная программа, отыскивающая в русскоязычном тексте количественные и порядковые числительные, выраженные как цифрами (числа), так и словесно в разных словоформах. Поиск основан на сличении слов текста со словарной базой из словаря: «М. Хаген – Полная парадигма. Морфология. Частотный словарь. Совмещенный словарь» (http://speakrus.ru/dict2/#morph-paradigm).

Предварительно из текста автоматически удалялись идиоматические выражения и устойчивые фразы, *случайно* содержащие числительные («семь пятниц на неделе», «ясно как дважды два четыре» и т.п.); вручную удалялись числительные, не связанные с авторским художественным замыслом – номера страниц, глав, перечисления 1), 2), 3), … и т.п.

Мы проанализировали наиболее объёмные произведения Бунина, Куприна, Сологуба и Арцыбашева, представленные в табл. 1.

Таблица 1 - Встречаемость числительных в исследованных произведениях

DOI: https://doi.org/10.18454/RULB.2023.45.28.1

№	Автор, текст, год создания	Объём (байты, кодировка UTF)	Количество числительных	Обратная плотность числительных
1	Бунин, *Жизнь Арсеньева* (1929)	987966	588	1680,2
2	Бунин, *Стихотворения*	921097	322	2860,5
3	Бунин, *Темные аллеи* (опубл. 1943)	796064	548	1452,7
4	Бунин, *Под серпом и молотом* (опубл. 1950)	756078	944	800,9
5	Бунин, *Деревня* (1910)	413138	272	1518,9
6	Бунин, *Окаянные дни* (1920)	408271	488	836,6
7	Куприн, *Яма* (1915)	1071674	1062	1009,1
8	Куприн, *Юнкера* (1932)	837555	992	844,3
9	Куприн, *Поединок* (1905)	795748	730	1090,1
10	Куприн, *Лазурные берега* (1913)	312916	492	636,0
11	Куприн, *Киевские типы* (1897)	166941	169	987,8
12	Сологуб, *Стихотворения*	1037506	357	2906,2

13	Сологуб, *Мелкий бес* (1902)	887247	393	2257,6
14	Арцыбашев, *У последней черты* (1912)	1566573	774	2024,0
15	Арцыбашев, *Санин* (1907)	1045996	427	2449,6
16	Арцыбашев, *Женщина, стоящая посреди* (1915)	409174	178	2298,7
17	Арцыбашев, *Смерть Ланде* (1904)	376880	149	2529,4

Основные результаты

Для каждого произведения найдена обратная плотность числительных как результат деления объёма текста на количество найденных в нём числительных. Чем меньше обратная плотность, тем чаще в тексте встречаются числительные.

В связи с этим понятны сравнительно малые значения обратной плотности в мемуарном (№4), дневниковом (№6) и очерковых (№10, 11) текстах, в которых неизбежно обилие фактографических числовых подробностей.

Сравнение обратных плотностей числительных для художественных текстов обнаруживает существенное различие между произведениями Бунина (№1, 3, 5) и Куприна (№7, 8, 9): в текстах последнего числительные используются *чаще* (детализация больше).

Наконец, обратим внимание на большие обратные плотности для поэзии (№2, 12), которой, вообще говоря, не свойственна детализация.

Проза Сологуба и Арцыбашева, добавленных в качестве impostors (№13–17), отличается очень большими значениями обратной плотности, статистически достоверно отличными от значений для прозы Бунина и Куприна.

Эти результаты показывают, что использование числительных специфично для автора и жанра.

В табл. 2 представлены абсолютные частоты числительных 1, 2, …, 5, которые содержатся во *всех* исследованных произведениях. Поскольку тексты сильно различаются по размеру (см. табл. 1), для сравнимости абсолютных частот числительных в разных текстах мы ввели поправочные коэффициенты, выбрав в качестве эталонного текста для сравнения «Яму» Куприна. Поэтому, например, частоты числительных в «Жизни Арсеньева» Бунина пришлось умножить на 1071674/987966 = 1,08, а для романа «У последней черты» Арцыбашева – на 1071674/1566573 = 0,68.

Абсолютные частоты числительных с поправкой на размер текста приведены в скобках в табл. 2.

Таблица 2 - Абсолютные частоты числительных 1, 2, …, 5 в исследованных текстах и исправленные абсолютные частоты (в скобках) с поправками, учитывающими разный размер текстов

DOI: https://doi.org/10.18454/RULB.2023.45.28.2

№	Автор, текст	числительные				
		1	2	3	4	5
1	Бунин, *Жизнь Арсеньева*	355 (385,1)	101 (109,6)	50 (54,2)	4 (4,3)	10 (10,9)
2	Бунин, *Стихотворения*	152 (176,9)	56 (65,2)	19 (22,1)	15 (17, 5)	6 (7,0)
3	Бунин, *Темные аллеи*	248 (333,9)	99 (133,3)	46 (61,9)	13 (17,5)	20 (26,9)
4	Бунин, *Под серпом и молотом*	306 (433,7)	99 (140,3)	48 (68,0)	14 (19,8)	18 (25,5)
5	Бунин, *Деревня*	105 (272,4)	46 (119,3)	31 (80,4)	7 (18,2)	13 (33,7)
6	Бунин, *Окаянные дни*	140 (367,5)	53 (139,1)	30 (78,8)	10 (26,3)	9 (23,6)
7	Куприн,	392 (392)	172 (172)	101 (101)	34 (34)	47 (47)

	Яма					
8	Куприн, *Юнкера*	313 (400,5)	208 (266,1)	112 (143,3)	80 (102,4)	27 (34,6)
9	Куприн, *Поединок*	302 (406,7)	127 (171,0)	66 (88,9)	25 (33,7)	38 (51,2)
10	Куприн, *Лазурные берега*	140 (479,5)	88 (301,4)	54 (184,9)	25 (85,6)	26 (89,0)
11	Куприн, *Киевские типы*	59 (378,8)	26 (166,9)	22 (141,2)	11 (70,6)	6 (38,5)
12	Сологуб, *Стихотворения*	209 (215,9)	72 (74,4)	25 (25,8)	12 (12,4)	3 (3,1)
13	Сологуб, *Мелкий бес*	176 (212,6)	79 (95,4)	50 (60,4)	15 (18,1)	12 (14,5)
14	Арцыбашев, *У последней черты*	496 (339,3)	105 (71,8)	46 (31,5)	9 (6,2)	7 (4,8)
15	Арцыбашев, *Санин*	283 (290,0)	72 (73,8)	22 (22,5)	1 (1,0)	3 (3,1)
16	Арцыбашев, *Женщина, стоящая посреди*	122 (319,5)	30 (78,6)	10 (26,2)	5 (13,1)	2 (5,2)
17	Арцыбашев, *Смерть Ланде*	106 (301,4)	22 (62,6)	4 (11,4)	2 (5,7)	3 (8,5)

Ещё более определённые результаты, чем при анализе обратных плотностей числительных, были получены при использовании иерархического кластерного анализа, объединяющего объекты в кластеры по принципу сходства. Как известно, мерой его является метрика ρ («расстояние»): чем меньше «расстояние» между объектами, тем больше сходство между ними. Мы применили манхэттенскую метрику

$$\rho(\boldsymbol{x}, \boldsymbol{y}) = \sum_i^n |x_i - y_i|, \quad (1)$$

где $\boldsymbol{x}$ и $\boldsymbol{y}$ – n-мерные векторы, компонентами которых являются исправленные абсолютные частоты (см. табл. 2) первых n натуральных чисел в двух анализируемых текстах (здесь n = 5, т. к. во всех исследованных текстах встречались числительные от одного до пяти).

В процессе кластеризации использован метод дальнего соседа, который приводит к образованию компактных, чётко очерченных кластеров [15], [16].

Как известно, выбор метрики и метода кластеризации невозможно строго обосновать; между тем, они способны существенно повлиять на результаты кластеризации. В нашем случае результаты оказались достаточно устойчивыми; другие разумные комбинации метрики и метода кластеризации лишь несущественно влияли на вид дендрограммы (рис. 1).

Исследованные тексты практически идеально распределились по кластерам в соответствии с авторством и жанром.

Использование числительных в текстах Бунина более единообразно, чем у Куприна: все прозаические тексты попадают в родственные кластеры с небольшой высотой слияния. Единственный файл, попавший в другой кластер – это стихи Бунина, которые вполне логично кластеризовались вместе с поэзией Сологуба.

Наблюдается некоторая временна́я эволюция использования числительных: произведения, близкие по времени создания, чаще попадают в один или родственные кластеры.

Отметим исключительную единообразность текстов Арцыбашева с точки зрения использования числительных: кластеры объединяются на небольшой высоте. Частичным объяснением этого, в свете отмеченной выше временно́й эволюции, может служить сравнительная непродолжительность творческого пути Арцыбашева.

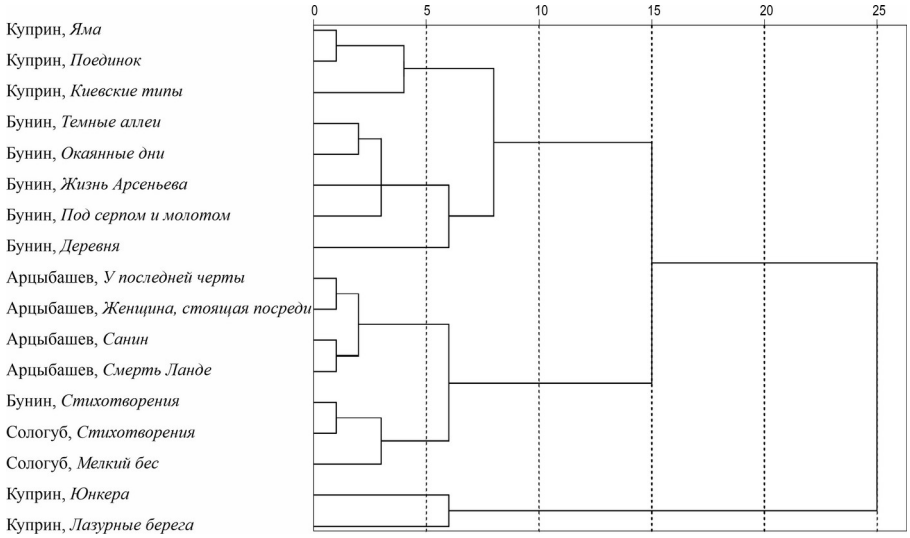

Рисунок 1 - Результат применения иерархического кластерного анализа к литературным текстам И.А. Бунина, А.И. Куприна, Ф.К. Сологуба и М.П. Арцыбашева. По вертикальной оси отложено «расстояние» в произвольных единицах
DOI: https://doi.org/10.18454/RULB.2023.45.28.3

Заключение

Разрабатываемый нами новый подход к задачам стилометрии, основанный на анализе статистики числительных в текстах, при всей его простоте, демонстрирует высокую эффективность и чувствительность. Показано, что манера использования числительных индивидуальна у каждого автора; их совокупность является характерным признаком, различающим тексты разных авторов. Тексты И.А. Бунина и А.И. Куприна, сравнительный анализ которых выполнялся до сих пор лишь в рамках традиционного описательного филологического подхода, впервые подвергнуты формальному количественному анализу, правильно распределившему тексты согласно авторству и жанрам. Использование числительных в текстах Бунина оказалось более единообразным, чем у Куприна. Привлечение для анализа текстов сторонних авторов (impostors) – Ф.К. Сологуба и М.П. Арцыбашева – усиливает значимость полученного результата и подтверждает его неслучайный характер.

Финансирование

Исследование выполнено за счет средств гранта Российского научного фонда № 23-28-00750, https://rscf.ru/project/23-28-00750/, проект «Разработка нового метода стилометрии на основе статистики использования числительных в авторских текстах».

Конфликт интересов

Не указан.

Рецензия

Сообщество рецензентов Международного научно-исследовательского журнала
DOI: https://doi.org/10.18454/RULB.2023.45.28.4

Funding

The research was supported by the grant No. 23-28-00750 from the Russian Science Foundation; see https://rscf.ru/en/project/23-28-00750/, Project "Development of a New Method of Stylometry Based on Statistics of Numerals Use in Authors' Texts".

Conflict of Interest

None declared.

Review

International Research Journal Reviewers Community
DOI: https://doi.org/10.18454/RULB.2023.45.28.4

Список литературы / References

1. Stamatatos E. A Survey of Modern Authorship Attribution Methods / E. Stamatatos // J. of the American Society for information Science and Technology. — 2009. — V. 60, No. 3. — P. 538–556.

2. Tempestt N. Stylometry Techniques and Applications / N. Tempestt, S. Kalaivani, F. Aneez [et al.] // ACM Comput. Surv. — 2017. — 50(6). — Article 86. — 36 p. — DOI: 10.1145/3132039.

3. Burrows J. Delta: a Measure of Stylistic Difference and a Guide to Likely Authorship / J. Burrows // Literary and Linguistic Computing. — 2002. — 17(3). — P. 267–287.

4. Brocardo M. L. Authorship Verification Using Deep Belief Network Systems / M. L. Brocardo, I. Traore, I. Woungang [et al.] // Int. J. Commun. Syst., 2017. — DOI: 10.1002/dac.3259.

5. Зенков А. В. Отклонения от закона Бенфорда и распознавание авторских особенностей в текстах / А. В. Зенков // Компьютерные исследования и моделирование. — 2015. — Т. 7. — № 1. — С. 197−201.

6. Зенков А. В. Новый метод стилеметрии на основе статистики числительных / А. В. Зенков // Компьютерные исследования и моделирование. — 2017. — Т. 9. — № 5. — С. 837–850.

7. Zenkov A. V. A Method of Text Attribution Based on the Statistics of Numerals / A. V. Zenkov // Journal of Quantitative Linguistics. — 2018. — V. 25. — No. 3. — P. 256–270. — DOI: 10.1080/09296174.2017.1371915.

8. Zenkov A. V. The Romantic Clash: Influence of Karel Sabina over Macha's Cikani from the Perspective of the Numerals Usage Statistics / A. V. Zenkov, M. Místecký // Glottometrics. — 2019. — V. 46. — P. 12–28.

9. Zenkov A. V. Stylometry and Numerals Usage: Benford's Law and Beyond / A. V. Zenkov // Stats. — 2021. — V. 4. — P. 1051–1068. — DOI: 10.3390/stats4040060.

10. Zenkov A. Young Vladimír Vašek? – A Numerals Analysis Contribution to the Bezruč−Hrzánský Identity Issue / A. Zenkov, M. Místecký // Naše řeč, 2022. — V. 105(3). — P. 151–161.

11. Бунин И. А. Pro et contra (Русский Путь) / И. А. Бунин. — СПб: Изд-во Русского Христианского гуманитарного института, 2001. — 1016 с. — ISBN 5-88812-066-9.

12. Гейдеко В. А. А. Чехов и Ив. Бунин: Монография / В.А. Гейдеко. — 2-е изд. — М.: Сов. писатель, 1987. — 368 с.

13. Смирнова Л. А. Иван Алексеевич Бунин: Жизнь и творчество / Л. А. Смирнова. — М.: Просвещение, 1991. — 192 с. — ISBN 5-09-002599-1.

14. Koppel M. Determining if Two Documents are Written by the Same Author / M. Koppel, Y. Winter // J. of the Association for Information Science and Technology. — 2014. — V. 65. — No. 1. — P. 178–187.

15. Gan Guojun. Data Clustering: Theory, Algorithms, and Applications, ASA-SIAM Series on Statistics and Applied Probability / Guojun Gan, Ma Chaoqun, Wu Jianhong // SIAM, Philadelphia, ASA, Alexandria, VA. — 2007. — 466 p. — ISBN 978-0-898716-23-8.

16. Moisl H. Cluster Analysis for Corpus Linguistics / H. Moisl. — Berlin, München, Boston: De Gruyter Mouton, 2015. — DOI: 10.1515/9783110363814.

Список литературы на английском языке / References in English

1. Stamatatos E. A Survey of Modern Authorship Attribution Methods / E. Stamatatos // J. of the American Society for information Science and Technology. — 2009. — V. 60, No. 3. — P. 538–556.

2. Tempestt N. Stylometry Techniques and Applications / N. Tempestt, S. Kalaivani, F. Aneez [et al.] // ACM Comput. Surv. — 2017. — 50(6). — Article 86. — 36 p. — DOI: 10.1145/3132039.

3. Burrows J. Delta: a Measure of Stylistic Difference and a Guide to Likely Authorship / J. Burrows // Literary and Linguistic Computing. — 2002. — 17(3). — P. 267–287.

4. Brocardo M. L. Authorship Verification Using Deep Belief Network Systems / M. L. Brocardo, I. Traore, I. Woungang [et al.] // Int. J. Commun. Syst., 2017. — DOI: 10.1002/dac.3259.

5. Zenkov A. V. Otkloneniya ot zakona Benforda i raspoznavanie avtorskih osobennostej v tekstah [Deviations from Benford's Law and Recognition of Copyright Features in Texts] / A. V. Zenkov // Komp'yuternye issledovaniya i modelirovanie [Computer Research and Modeling]. — 2015. — V. 7. — № 1. — P. 197−201 [in Russian].

6. Zenkov A. V. Novyj metod stilemetrii na osnove statistiki chislitel'nyh [A New Method of Telemetry Based on Numerals Statistics] / A. V. Zenkov // Komp'yuternye issledovaniya i modelirovanie [Computer Research and Modeling]. — 2017. — V. 9. — № 5. — P. 837–850 [in Russian].

7. Zenkov A. V. A Method of Text Attribution Based on the Statistics of Numerals / A. V. Zenkov // Journal of Quantitative Linguistics. — 2018. — V. 25. — No. 3. — P. 256–270. — DOI: 10.1080/09296174.2017.1371915.

8. Zenkov A. V. The Romantic Clash: Influence of Karel Sabina over Macha's Cikani from the Perspective of the Numerals Usage Statistics / A. V. Zenkov, M. Místecký // Glottometrics. — 2019. — V. 46. — P. 12–28.

9. Zenkov A. V. Stylometry and Numerals Usage: Benford's Law and Beyond / A. V. Zenkov // Stats. — 2021. — V. 4. — P. 1051–1068. — DOI: 10.3390/stats4040060.

10. Zenkov A. Young Vladimír Vašek? – A Numerals Analysis Contribution to the Bezruč−Hrzánský Identity Issue / A. Zenkov, M. Místecký // Naše řeč, 2022. — V. 105(3). — P. 151–161.

11. Bunin I. A. Pro et contra (Russkij Put') [Pro et contra (The Russian Way')] / I. A. Bunin. — SPb: Publishing House of the Russian Christian Humanitarian Institute, 2001. — 1016 p. — ISBN 5-88812-066-9 [in Russian].

12. Gejdeko V. A. A. CHekhov i Iv. Bunin: Monografiya [A. Chekhov and I.V. Bunin: Monograph] / V.A. Gejdeko. — 2nd ed. — M.: Soviet Writer, 1987. — 368 p. [in Russian]

13. Smirnova L. A. Ivan Alekseevich Bunin: ZHizn' i tvorchestvo [Ivan Alekseevich Bunin: Life and Work] / L. A. Smirnova. — M.: Prosveshchenie, 1991. — 192 p. — ISBN 5-09-002599-1 [in Russian].

14. Koppel M. Determining if Two Documents are Written by the Same Author / M. Koppel, Y. Winter // J. of the Association for Information Science and Technology. — 2014. — V. 65. — No. 1. — P. 178–187.

15. Gan Guojun. Data Clustering: Theory, Algorithms, and Applications, ASA-SIAM Series on Statistics and Applied Probability / Guojun Gan, Ma Chaoqun, Wu Jianhong // SIAM, Philadelphia, ASA, Alexandria, VA. — 2007. — 466 p. — ISBN 978-0-898716-23-8.

16. Moisl H. Cluster Analysis for Corpus Linguistics / H. Moisl. — Berlin, München, Boston: De Gruyter Mouton, 2015. — DOI: 10.1515/9783110363814.