РУССКИЙ ЯЗЫК, ЯЗЫКИ НАРОДОВ РОССИИ. ТЕОРЕТИЧЕСКАЯ И ПРИКЛАДНАЯ ЛИНГВИСТИКА

УДК 81-13:81’32:811.512.145
DOI 10.22213/2618-9763-2024-2-107-117

М. А. Комышев, студент
В. А. Баранов, доктор филологических наук, профессор
Ижевский государственный технический университет имени М. Т. Калашникова, Ижевск, Россия

СТИЛОМЕТРИЧЕСКИЙ АНАЛИЗ ПРОИЗВЕДЕНИЙ ГАБДУЛЛЫ ТУКАЯ И САГИТА СУНЧЕЛЕЯ: КТО АВТОР СТИХОТВОРЕНИЯ «СӨЙ ГОМЕРНЕ…»?

Данная работа посвящена определению авторства татароязычного четверостишия «Сөй гомерне…» («Люби жизнь…») с помощью методов стилометрии. В качестве возможных авторов рассматриваются Габдулла Тукай, кому традиционно приписывался этот текст, и Сагит Сунчелей, свидетельства в пользу авторства которого были обнаружены литературоведами на рубеже XX–XXI веков. В работе проводится апробация нескольких методов стилометрии на основе расстояния между текстами.

Материалом исследования стал корпус оцифрованных оригинальных стихотворных произведений Габдуллы Тукая и Сагита Сунчелея на современном литературном варианте татарского языка. В итоговую выборку для анализа были включены 10 текстов Тукая и 8 текстов Сунчелея. Измерено расстояние между текстами посредством метрик Delta Бёрроуза, Cosine Delta и косинусного расстояния. С помощью пакета Stylo выполнены вычисления и на их основе построены дендрограммы.

Наиболее эффективной показала себя мера Cosine Delta с извлечением символьных n-грамм (при n = 4). Это объясняется, во-первых, агглютинативным строем татарского языка, т. к. среди наиболее частотных извлекаемых n-грамм обнаруживаются элементы, соответствующие формообразующим аффиксам. Во-вторых, использование символьных n-грамм позволяет увеличить объем анализируемых данных. В некоторых случаях качественной атрибуции удалось добиться также с помощью Cosine Delta на основе словоформ и косинусного расстояния на основе символьных n-грамм.

Несмотря на невозможность сделать однозначный вывод из-за крайне малых объемов анализируемых текстов, сопоставление результатов применения различных методов показывает, что наиболее вероятным автором проблемного четверостишия является Сагит Сунчелей.

Ключевые слова: авторство; стилометрия; метод *Delta* Бёрроуза; Габдулла Тукай; Сагит Сунчелей; «Сөй гомерне…»; татарский язык.

Введение

Задача определения автора анонимного, подложного или спорного текста является одной из традиционных в области литературоведения, археографии, дипломатики, истории, судебного авторoведения, лингвистики и некоторых других теоретических и прикладных направлений, объектом которых является текст, произведение. В настоящее время методы и инструменты для решения этой задачи предлагает стилометрия – междисциплинарное направление, изучающее индивидуальный стиль автора (идиостиль) или стиль отдельных текстов как набор неких численно измеримых параметров. В основе данного подхода лежит явление, заключающееся в том, что каждый бессознательно предпочитает использовать в речи конкретные языковые единицы и конструкции, в совокупности представляющие собой его уникальную речевую характеристику; эти особенности употребления языковых средств и являются объектом и предметом анализа стилометрии [1–2].

Развитие стилометрии как направления и ее основные методы подробно рассмотре-

ны в обзорных работах Г. Я. Мартынова и других авторов (например, [3–6]). В них описаны различные извлекаемые из текстов лексические и структурные маркеры (наиболее частотные слова, длина слов или предложений, богатство лексикона и т. п.), а также методы сопоставления и классификации текстов на их основе (с применением технологий машинного обучения, вычислением расстояний между текстами и др.). Огромное количество работ посвящено использованию стилометрических методов для определения авторства текстов, написанных на многих европейских языках – английском [7], испанском [8], русском [9] и других. В то же время применение стилометрии в отношении текстов на языках народов России только начинается. Как отмечают А. М. Галиева и А. А. Аминова, «…в настоящее время стилометрия художественных текстов на татарском языке только формируется…» [10].

Материалом для данного исследования стали корпусы стихотворных текстов Габдуллы Тукая и Сагита Сунчелея. Габдулла Тукай (1886–1913) – поэт-классик татарской литературы начала XX века. Сагит Сунчелей (1888–1937) – татарский поэт, драматург, переводчик, чье творчество также пришлось на упомянутый период и который находился в тесном общении с Тукаем. Сунчелей в 1930-х гг. был репрессирован по обвинению в «контрреволюционной султангалиевщине», вследствие чего его творчество на долгие годы оставалось недоступно как читателям, так и исследователям. До сих пор задача изучения творчества и личности Сагита Сунчелея остается актуальной для татарского литературоведения [11].

Данная работа посвящена определению авторства четверостишия «Сөй гомерне…» («Люби жизнь…»), традиционно приписываемого Тукаю:

Сөй гомерне, сөй халыкны, сөй халыкның дөньясын,

Люби жизнь, люби народ, люби мир народа (то, чем живёт народ),

Без үләрбез, билгеле, тик үкенечкә калмасын.

Мы умрём, разумеется, пусть это не огорчает.

Киң күңелле бул эченнән, мыскыл ишетсәң, «вак» диген,

Будь широким душой, если слышишь оскорбления, говори: «Мелочность»,

Таптасыннар, хурласыннар, тик җаның хурланмасын.

Пусть топчут, пусть унижают – пусть твоя душа совсем не будет посрамлена.

Как отмечает литературовед З. З. Рамеев [12], данное четверостишие включается в многотомные собрания сочинений Г. Тукая с 40-х годов XX века. Одновременно З. З. Рамеев показал, что настоящим автором этого стихотворения следует считать Сагита Сунчелея: четверостишие в действительности является частью более крупного текста «Яшә!» («Живи!»), опубликованного в 1912 г. в журнале «Йолдыз» под его именем. Там оно представлено в несколько ином виде: в частности, на месте слов «халыкның дөньясын» находится «ходаның дөньясын». Несмотря на то что выводы З. З. Рамеева не вызывают сомнений, до сих пор во многих источниках – как печатных, так и электронных – авторство данного текста приписывается Габдулле Тукаю.

Целью работы является нахождение на основе стилометрических методов дополнительных доказательств принадлежности спорного четверостишья перу или Габдуллы Тукая, или Сагита Сунчелея.

Задачи:

а) апробация некоторых методов стилометрии на основе расстояний между текстами для определения авторства стихотворных произведений на татарском языке;

б) определение с помощью статистических методов наиболее вероятного автора четверостишия «Сөй гомерне…».

Тюркские языки как объект стилометрического анализа

Татарский язык – язык кыпчакской группы тюркской ветви алтайской семьи языков, государственный язык Республики Татарстан [13]. Имеет агглютинативный строй, подобно родственному турецкому, а также, например, финно-угорским языкам – финскому, удмуртскому и некоторым другим. Как и в других агглютинативных языках, в татарском языке словоформы, указывающие на грамматические связи в предложении, образуются с помощью аффиксов, которые последовательно добавляются к концам слов [14].

Ввиду отсутствия стилометрических исследований текстов на татарском языке[1], мы можем обратиться к работам, посвященным текстам на родственном турецком языке.

Результаты анализа турецкоязычных текстов с помощью стилометрических методов представлены, например, в [16–19]: Ф. Джан, Дж. Паттон (*F. Can*, *J. M. Patton*) предлагают в качестве стилистического маркера, дающего вполне хорошие результаты, использовать длину слов, т. к. «[а]вторы имеют больше возможностей (бессознательно) контролировать длину слов в агглютинативных языках вроде турецкого». Вероятно, в поэтических текстах этот параметр менее показателен, поскольку выбор словоформ ограничивается еще и законами стихосложения. Там же утверждается, что точность классификации снижается при отсечении аффиксов, поэтому делается вывод, что они представляют важную стилистическую информацию; Х. Такчи, Э. Экинчи (H. Takçı, E. Ekinci) рекомендуют использование отдельных единичных символов (заглавные и строчные буквы алфавита, знаки пунктуации, пробелы, символы переноса строки) как характеристику, позволяющую достаточно успешно производить атрибуцию текстов (в исследовании на материале статей новостного сайта *Sabah* получен средний уровень достоверности в 86 %).

Delta Бёрроуза и другие методы стилометрии

Одним из методов автоматической атрибуции текста является метод *Delta* Джона Бёрроуза, представленный в 2001 г. [20]. Как утверждается в [21], данный метод стал «широко используемым и общепринятым методом» в стилометрии.

Начальным этапом применения данного метода является преобразование текстов в представление, именуемое «мешком слов» (*bagofwords*): подсчитывается, сколько раз каждое слово встречается в каждом анализируемом тексте. Эти абсолютные количества вхождений преобразуются в относительные частоты, и для анализа из всего набора (корпуса) текстов отбираются N наиболее часто встречающихся слов $\{w_i\}_{i=1}^{N}$. Обозначив относительную частоту слова w_i в тексте D как $f_i(D)$, каждый текст можно представить в виде вектора $\vec{f}(D)=\left(f_1(D), f_2(D), \ldots, f_N(D)\right)$, где каждое слово имеет свое измерение. Под «словами» мы можем понимать различные объекты: непосредственно словоформы, леммы, символы или их *n*-граммы – последовательности из $n \in \mathbb{N}$ идущих подряд элементов (при $n=1$ мы выделяем отдельные слова, символы и т. д., их будем называть *униграммами*).

Метод *Delta* Дж. Бёрроуза (или, как он иначе называется, *Classic Delta*) определяет расстояние между двумя текстами D и D' как «среднее арифметическое абсолютных разностей *z*-оценок для множества слов-переменных…» [22]:

$$\Delta(D, D') = \frac{1}{N}\sum_{i=1}^{N}\left|z\left(f_i(D)\right) - z\left(f_i(D')\right)\right|. \quad (1)$$

В то же время *z*-оценка для слова w_i в документе D определяется следующим образом:

$$z\left(f_i(D)\right) = \frac{f_i(D) - \mu_i}{\sigma_i}, \quad (2)$$

[1] В работе [15], посвященной исследованию текстов на татарском языке традиционными методами, атрибуция авторства с помощью статистических приемов лишь упоминается.

где μ_i – средняя частота данного слова в анализируемом корпусе, σ_i – среднеквадратичное отклонение частоты.

Имеются различные вариации метода *Delta* Дж. Бёрроуза: *Cosine Delta*, *Eder* модификация *Cosine Delta*, где манхэттенское расстояние в (1) заменено на косинусное расстояние.

Метод *Delta*, согласно исследованиям [24], показывает себя наиболее эффективным при анализе текстов на английском языке – достоверность атрибуции текстов на других языках ниже, но тем не менее приемлема. Как пример использования можем указать на одно из относительно недавних крупных стилометрических исследований на материале русского языка – работу Н. П. Великановой и Б. В. Орехова по определению авторства романа «Тихий Дон» [25]. Авторами было показано, что статистически наиболее вероятным автором романа является сам М. А. Шолохов.

В этой же работе авторы утверждают, что «…надежные результаты [с помощью метода *Delta*] могут быть получены только на жанрово гомогенном корпусе, а каждый составляющий этот корпус текст должен быть по объему не меньше 10 000 слов». Однако некоторые выводы могут быть сделаны и при использовании корпуса меньшего объема, особенно когда производится сопоставление с результатами применения других методов[1].

Помимо *Delta* Дж. Бёрроуза, для вычисления расстояния между текстами могут быть применены и другие метрики. В частности, в данном исследовании мы рассмотрим использование «классического» для векторной алгебры косинусного расстояния:

$$d(D,D') = 1 - \frac{\vec{f}(D)\cdot\vec{f}(D')}{\left|\vec{f}(D)\right|\left|\vec{f}(D')\right|}. \qquad (3)$$

где $\vec{f}(D)$ и $\vec{f}(D')$ – векторы текстов, полученные описанным выше образом.

Для проведения настоящего исследования был использован пакет *Stylo* [27] для среды программирования *R*. Этот пакет реализует метод *Delta* и его модификации, некоторые другие меры расстояния, включая *Delta*, *Argamon Delta* и др. Исследования, например [23], показывают, что наиболее высокое качество атрибуции (по крайней мере, на корпусах романов на немецком, английском и французском языках) дает косинусное расстояние, различные способы извлечения слов, символов и их *n*-грамм, а также возможности проведения кластеризационного анализа текстов и визуализации результатов в виде дендрограмм. Дендрограмма представляет собой древовидный граф, «листьями» которого выступают тексты, и те из них, которые более близки по стилистике, размещаются на нем рядом в одной «ветви»-кластере.

Текстовый материал

Для проведения исследования был собран корпус стихотворных произведений Габдуллы Тукая и Сагита Сунчелея в электронной форме. При подготовке корпуса отдавалось предпочтение текстам, объем которых составляет более 100 словоупотреблений и которые близки по жанровой форме к проблемному тексту. Не включались в корпус тексты, которые, согласно комментариям в источнике (авторским или составителя), были вдохновлены другими произведениями или заимствованы полностью или частично. Проводились эксперименты с различными составами корпуса, в том числе с добавлением жанрово гетерогенных произведений.

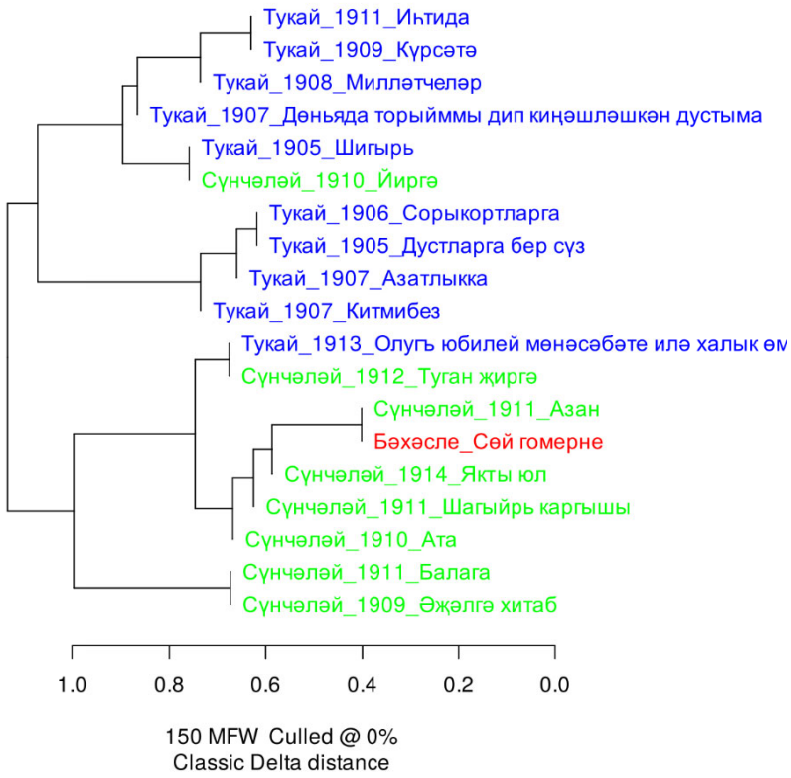
Тексты Габдуллы Тукая на современном литературном варианте татарского языка были взяты с веб-сайта «Тукай дөньясы»[2]. Машиночитаемые варианты текстов Сагита Сунчелея были вручную подготовлены на основе собрания сочинений [28]. В тех случаях, когда это потребовалось, тексты были очищены от сносок, пунктуация была приведена к единому виду, слова с переносами переведены в одну строку. Мы также снабдили тексты информацией о годе публикации или – если это указано в источнике – годе написания.

Всего в ходе работы было собрано 28 текстов Тукая и столько же текстов Сунчелея. В итоговую выборку для анализа, исходя из требования жанровой гомогенности были включены 10 текстов Тукая и 8 текстов,

Сунчелея. Общее количество словоупотреблений в итоговой выборке, включая спорный текст, составило 3140, в среднем на текст – 165 словоформ.

Результаты анализа и их обсуждение

Метод *Classic Delta* с униграммами слов показал себя не самым эффективным. Несмотря на то что можно проследить некоторые закономерности в кластеризации (например, тексты, написанные примерно в одно время, имеют тенденцию размещаться рядом друг с другом), нам не удалось получить качественное разделение произведений двух авторов (рис. 1). Изменение параметров метода (количество слов, процент выбраковки[1]) позволяет получать более ясные результаты.



Источник: выполнен авторами.

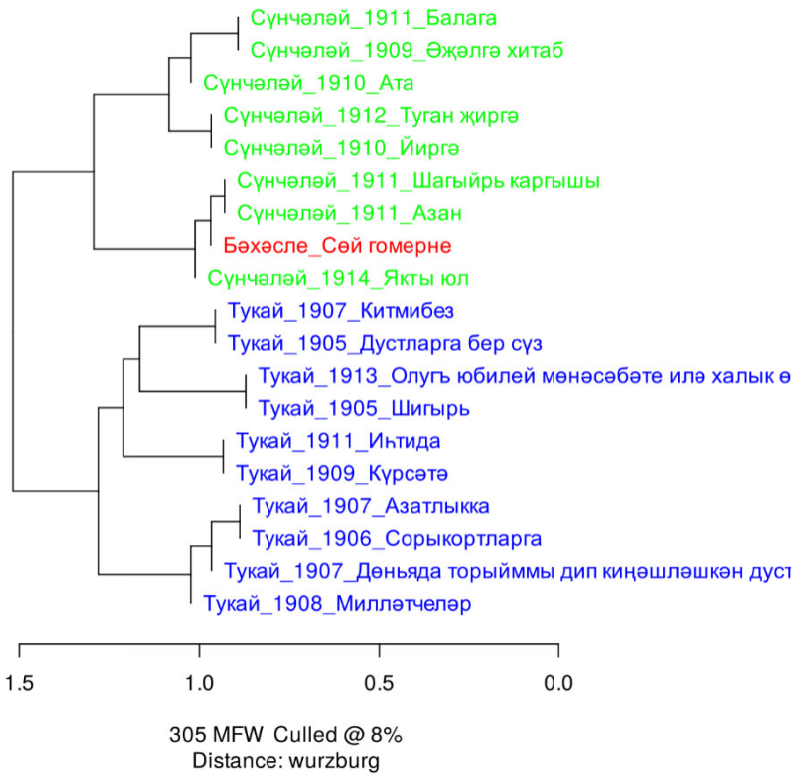
Рис. 1. Результаты использования меры *Classic Delta* с униграммами слов

Fig. 1. Results of using the Classic Delta measure with word unigrams

Использование меры *Cosine Delta* вместе с униграммами слов позволяет достичь значительного улучшения кластеризации; в некоторых случаях мы можем путем подбора числа слов и процента выбраковки получить полностью корректную атрибуцию.

Наиболее эффективным для сравнения векторов текстов показало себя использование символьных *n*-грамм (приемлемые результаты можно получить при $n = 3,4$, более высокое качество кластеризации мы получили при $n = 4$) и *Cosine Delta*: подобрав нужный процент выбраковки, удалось добиться разделения текстов Тукая и Сунчелея по отдельным веткам (рис. 2).

[1] Процент выбраковки (*culling*) указывает минимальный процент текстов корпуса, в которых должно встретиться извлекаемое слово для того, чтобы учитываться при вычислении расстояния. В данном исследовании в большинстве случаев было необходимо указать некий процент, чтобы получить удовлетворительные результаты.



Источник: выполнен авторами

Рис. 2. Результаты использования меры *Cosine Delta* с символьными 4-граммами

Fig. 2. Results of using the Cosine Delta measure with character 4-grams

Эффективность такого подхода можно объяснить, *во-первых*, агглютинативным строем татарского языка. Это подтверждается, например, списком 4-грамм, которые извлекает *Stylo* (таблица), в котором, помимо небольших служебных слов и местоимений, а также, вероятно, начальных элементов некоторых знаменательных слов, представлено и некоторое количество аффиксов – например, твердые и мягкие вариации аффиксов множественного числа (*-лар*, *-ләр*) и родительного падежа (*-ның*, *-нең*). Как показали стилометрические исследования текстов на турецком языке, очень важную информацию несет в себе форма слова, и использование *n*-грамм позволило нам выделить в том числе значимые компоненты слов – формообразующие аффиксы. *Во-вторых*, используя символьные *n*-граммы вместо целых слов, мы можем увеличить объем анализируемых данных, т. к. одно слово может производить несколько *n*-грамм; при исследовании малых по объему текстов это, как показал анализ, оказалось результативным.

25 наиболее частотных 4-грамм, извлекаемых *Stylo*

The 25 most frequent character 4-grams extracted by Stylo

1	_бер	6	бер_	11	*нең_*	16	*ләр_*	21	*сын_*
2	_мин	7	_дә_	12	_бар	17	*ргә_*	22	_кил
3	_бул	8	_син	13	_без	18	_җан	23	_кар
4	мин_	9	_ул_	14	_тор	19	*ның_*	24	_бел
5	*лар_*	10	*нда_*	15	*ган_*	20	_да_	25	_күр

Примечание. Знаком «_» в таблице обозначены символы пробела. Курсивом выделены 4-граммы, соответствующие формообразующим аффиксам.

Источник: составлена авторами.

Сходные результаты дает и использование «классического» косинусного расстояния (3) вместо *Cosine Delta*. Однако эти две метрики могут приводить к различной кластеризации. Нам удалось «ввести в заблуждение» косинусное расстояние путем добавления некоторых конкретных текстов, незначительно отличающихся от всех остальных, а иногда и наоборот – «слишком» похожих: например, при использовании косинусного расстояния стихотворение Тукая «Күрсәтә» (1909) устойчиво располагается рядом с «Сөй гомерне…» в ветке Сунчелея, что, возможно, объясняется схожестью этих двух текстов идейной направленностью, а соответственно, использованием аналогичных языковых средств[1]. *Cosine Delta* дает ошибочные результаты в основном при наличии в корпусе жанрово отличных текстов.

Отвечая на вопрос, почему *Cosine Delta* оказывается эффективнее, можно обратиться к упомянутому ранее исследованию вариаций метода *Delta* [29]. Используя *Cosine Delta* со стандартизацией частот по формуле (2), для каждого текста мы получаем в сущности тернарный «ключевой профиль» особенностей употребления языковых средств («чаще», «реже», чем в среднем по корпусу, или «примерно одинаково»), который, как показывают авторы исследования, оказывается достаточно эффективен в определении авторского стиля. Сочетание с извлечением *n*-грамм, вероятно, и помогло нам дополнительно повысить точность.

При использовании *Cosine Delta* с символьными *n*-граммами в тех случаях, когда все остальные тексты кластеризуются корректно, четверостишие «Сөй гомерне…» оказывается в ветке с текстами Сунчелея. При применении «классического» косинусного расстояния в большинстве случаев происходит аналогичная группировка. В то же время эксперименты с добавлением отдельных текстов показали диаметрально противоположные результаты, при которых «Сөй гомерне…» располагается в ветке Тукая, в то время как все остальные тексты кластеризуются корректно[2]. Необходим дополнительный анализ этих пограничных эффектов для того, чтобы можно было сделать однозначные выводы.

Выводы

В исследовании для определения авторства четверостишия «Сөй гомерне…», написанного на татарском языке, мы использовали несколько стилометрических методов.

Наиболее качественная атрибуция была получена при использовании символьных 4-грамм и меры *Cosine Delta*. Скорее всего, это связано с агглютинативной природой татарского языка. Поэтому стоит ожидать, что при автороведческом анализе проблемных текстов на татарском и других агглютинативных языках (например, удмуртском) такой метод также будет показывать высокую эффективность.

Заметим, что объем спорного стихотворения «Сөй гомерне…» крайне мал – 26 словоупотреблений. При анализе текстов малого объема методы стилометрии не претендуют на высокую достоверность, и в данной работе мы показали некоторые случаи, когда они дают некачественные результаты. Несмотря на это сопоставление результатов, полученных различными методами, позволяет утверждать, что статистически наиболее вероятным автором четверостишия является Сагит Сунчелей.

Библиографические ссылки

1. Understanding and explaining Delta measures for authorship attribution / Stefan Evert [et al.] // Digital Scholarship in the Humanities. 2017. Vol. 32. Suppl. 2. Pp. ii4–ii16. DOI: 10.1093/llc/fqx023

2. Криминалистика. Теоретический курс : монография / Ф. Г. Аминев [и др.]. Уфа : НИИ ППГ, 2022. С. 338–382. ISBN 978-5-91144-017-6. EDN: YZRDPV

[1] И в том, и в другом тексте автор призывает читателя жить с гордостью, показывая добродетельные качества.

[2] В частности, в некоторых случаях к такому эффекту приводило приобщение к корпусу текста Сунчелея «Тормышым» (1910). Об обоснованности его добавления можно судить скорее отрицательно: в то время как «Сөй гомерне…» вдохновляет читателя, рекомендует определенный стиль жизни, «Тормышым» посвящено размышлениям лирического героя о поиске лучшей жизни – соответственно, при его добавлении в корпус могут начать проявляться сторонние факторы в виде особенностей жанровой формы. Однако такое поведение было зафиксировано не при всех составах корпуса.

3. *Мартыненко Г. Я.* Стилеметрия: возникновение и становление в контексте междисциплинарного взаимодействия. Ч. 1. Первые шаги: XIX век // Структурная и прикладная лингвистика. 2014. Вып. 10. С. 3–23. EDN: WGSGST

4. *Мартыненко Г. Я.* Стилеметрия: возникновение и становление в контексте междисциплинарного взаимодействия. Ч. 2. Первая половина XX века // Структурная и прикладная лингвистика. 2015. Вып. 11. С. 10–28. EDN: VTRBAV

5. *Stamatatos E.* A Survey of Modern Authorship Attribution Methods // Journal of the American Society for Information Science and Technology. 2009. Vol. 60. Issue 3. Pp. 538–556. DOI: 10.1002/asi.21001

6. *MisiniA., Kadriu A., Canhasi E.* A Survey on Authorship Analysis Tasks and Techniques // SEEU Review. 2022. Vol. 17. Issue 2. Pp. 153–167. DOI: 10.2478/seeur-2022-0100

7. *Modrall Sperling D.H., Kestemont M., Neyt V.* The Authorship of Stephen King's Books Written Under the Pseudonym "Richard Bachman": A Stylometric Analysis // Journal of Computational Literary Studies. 2023. Vol. 2. Issue 1. Pp. 1–35. DOI: 10.48694/jcls.3594.

8. *CuéllarÁ., Vega García-Luengos G.* Un nuevo repertorio dramático para Andrés de Claramonte // Hipogrifo. 2023. Vol.11. Núm. 1. Pp. 117–172. DOI: 10.13035/H.2023.11.01.09

9. *Великанова Н. П.*, *Орехов Б. В.* Цифровая текстология: атрибуция текста на примере романа М. А. Шолохова «Тихий Дон» // Мир Шолохова. 2019. №1(11). С. 70–82. EDN: HQDAHD

10. *Галиева А. М.*, *Аминова А. А.* Метафорические образы времени в поэтическом творчестве Сулеймана // Филология и культура. 2020. № 2 (60). С. 18–26. DOI: 10.26907/2074-0239-2020-60-2-18-26. EDN: HJKHAX

11. *Сүнчәләй С.* Әсәрләр һәм хатлар / ТР ФА Г. Ибраһимов ис. Тел, әдәбият һәм сәнгать институты. Казан: Татарстан китап нәшрияты, 2005. 367 б. ISBN 5-298-04071-3

12. *Рәмиев З.* Тукай һәм замандаш әдипләр : «Габдулла Тукай» энциклопедик сүзлек-белешмәсенә материаллар / ТР ФА Гуманитар ф. бүлеме. Казан: ТаРИХ, 2004. 111б. ISBN 5-94113-099-6

13. Татар теле: үткәне һәм бүгенгесе / ТР ФА Г. Ибраһимов ис. Тел, әдәбият һәм сәнгать институты ; хезмәтне әзерләүчеләр: Р. Г. Галиуллин, Г. К. Һадиева. Казан, 2021. 72 б. ISBN 978-5-93091-369-9. EDN: ICNLSQ

14. Татар грамматикасы : өч томда / ТР ФА Г. Ибраһимов ис. Тел, әдәбият һәм сәнгать институты ; проект җит. М. З. Зәкиев ; ред. Ф. М. Хисамова. Тулыландырылган 2нче басма. Казан, 2015–2017. Т. 1. 2015. 512б. ISBN 978-5-93091-192-3. EDN: YUXEXM

15. *Галиева А. М.*, *Аминова А. А.* Метафорические образы времени в поэтическом творчестве Сулеймана // Филология и культура. 2020. № 2 (60). С. 18–26. DOI: 10.26907/2074-0239-2020-60-2-18-26. EDN: HJKHAX

16. *Can F.*, *Patton J. M.* Change of Writing Style with Time // Computers and the Humanities. 2004. Vol. 38. Pp. 61-82. DOI: 10.1023/B:CHUM.0000009225.28847.77

17. *Takçı H.*, *Ekinci E.* Character Level Authorship Attribution for Turkish Text Documents // The Online Journal of Science and Technology. 2012. Vol. 2. Issue 3. Pp. 12–16.

18. *Yülüce İ.*, *Dalkılıç F.* Author Identification with Machine Learning Algorithms // International Journal of Multidisciplinary Studies and Innovative Technologies. 2022. Vol. 6. Issue 1. Pp. 45–50. DOI: 10.36287/ijmsit.6.1.45

19. *Kocagül Yüzer H.* Authorship Attribution in Turkish Texts. Ankara: Artsürem, 2022. 220 p. ISBN 978-605-72285-0-5

20. *Burrows J.* "Delta": a Measure of Stylistic Difference and a Guide to Likely Authorship // Literary and Linguistic Computing. 2002. Vol. 17. Issue 3. Pp. 267–287. DOI: 10.1093/llc/17.3.267

21. Understanding and explaining Delta measures for authorship attribution / Stefan Evert [et al.] // Digital Scholarship in the Humanities. 2017. Vol. 32. Suppl. 2. Pp. ii4–ii16. DOI: 10.1093/llc/fqx023

22. *Argamon S.* Interpreting Burrow's Delta: Geometric and Probabilistic Foundations // Literary and Linguistic Computing. 2008. Vol. 23. Issue 2. Pp. 131–147. DOI: 10.1093/llc/fqn003

23. Understanding and explaining Delta measures for authorship attribution / Stefan Evert [et al.] // Digital Scholarship in the Humanities. 2017. Vol. 32. Suppl. 2. Pp. ii4–ii16. DOI: 10.1093/llc/fqx023

24. *Rybicki J.*, *Eder M.* Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words? // Literary and Linguistic Computing. 2010. Vol. 26. Issue 3. Pp. 315–321. DOI: 10.1093/llc/fqr031

25. *Великанова Н. П.*, *Орехов Б. В.* Цифровая текстология: атрибуция текста на примере романа М. А. Шолохова «Тихий Дон» // Мир Шолохова. 2019. № 1 (11). С. 70–82. EDN: HQDAHD

26. *Алиева О. В.* Delta Берроуза для древнегреческих авторов: опыт применения // Schole. 2022. Т. 16. Вып. 2. С. 693–705. DOI: 10.25205/1995-4328-2022-16-2-693-705. EDN: SIQWWO

27. *Eder M.*, *Rybicki J.*, *Kestemont M.* Stylometry with R: A Package for Computational Text Analysis // The R Journal. 2016. Vol. 8. No. 1. Pp. 107–121. DOI: 10.32614/RJ-2016-007

28. *Сүнчәләй С.* Әсәрләр һәм хатлар / ТР ФА Г. Ибраһимов ис. Тел, әдәбият һәм сәнгать институты. Казан: Татарстан китап нәшрияты, 2005. 367 б. ISBN 5-298-04071-3

29. Understanding and explaining Delta measures for authorship attribution / Stefan Evert [et al.] // Digital Scholarship in the Humanities. 2017. Vol. 32. Suppl. 2. Pp. ii4–ii16. DOI: 10.1093/llc/fqx023

References

1. Stefan Evert [et al.]. Understanding and explaining Delta measures for authorship attribution. Digital Scholarship in the Humanities, 2017, vol. 32, suppl. 2, pp. ii4-ii16. DOI: 10.1093/llc/fqx023

2. Aminev F.G. Jeksarhopulo A.A., Makarenko I.A., Zajnullin R.I. [i dr.]. *Kriminalistika. Teoreticheskij kurs : monografija* [Forensics. Theoretical course, monograph]. Ufa, NII PPG, 2022, pp. 338-382. (in Russ.). ISBN 978-5-91144-017-6. EDN: YZRDPV

3. Martynenko G.Ja. [Stylemetry: emergence and formation in the context of interdisciplinary interaction. Part 1. First steps: XIX century]. *Strukturnaja i prikladnaja lingvistika*, 2014, issue 10, pp. 3-23. (in Russ.). EDN: WGSGST

4. Martynenko G.Ja. [Stylemetry: emergence and formation in the context of interdisciplinary interaction. Part 2. First steps: XIX century]. *Strukturnaja i prikladnaja lingvistika*, 2015, issue 11, pp. 10-28. (in Russ.). EDN: VTRBAV

5. Stamatatos E. A Survey of Modern Authorship Attribution Methods. Journal of the American Society for Information Science and Technology, 2009, vol. 60, issue 3, pp. 538-556. DOI: 10.1002/asi.21001

6. Misini A., Kadriu A., Canhasi E. A Survey on Authorship Analysis Tasks and Techniques. SEEU Review, 2022, vol. 17, issue 2, pp. 153-167. DOI: 10.2478/seeur-2022-0100

7. Modrall Sperling D.H., Kestemont M., Neyt V. The Authorship of Stephen King's Books Written Under the Pseudonym "Richard Bachman": A Stylometric Analysis. Journal of Computational Literary Studies, 2023, vol. 2, issue 1, pp. 1-35. DOI: 10.48694/jcls.3594

8. CuéllarÁ., Vega García-LuengosG. Un nuevo repertorio dramático para Andrés de Claramonte. Hipogrifo, 2023, vol. 11, núm. 1, pp. 117-172. DOI: 10.13035/H.2023.11.01.09

9. Velikanova N.P., Orehov B.V. [Digital textual criticism: text attribution using the example of M. A. Sholokhov's novel "Quiet Don"]. *Mir Sholohova*, 2019, no. 1 (11), pp. 70-82. (in Russ.). EDN: HQDAHD

10. Galieva A.M., Aminova A.A. [Metaphorical images of time in the poetic works of Su-leyman]. *Filologija i kul'tura*, 2020, no. 2 (60), pp. 18-26. (in Russ.). DOI: 10.26907/2074-0239-2020-60-2-18-26. EDN: HJKHAX

11. Сүнчәләй С. Әсәрләр һәм хатлар. ТР ФА Г. Ибраһимов ис. Тел, әдәбият һәм сәнгать институты. Казан : Татарстан китап нәшрияты, 2005, 367 б. ISBN 5-298-04071-3

12. Рәмиев З. Тукай һәм замандаш әдипләр : «Габдулла Тукай» энциклопедик сүзлек-белешмәсенә материаллар / ТР ФА Гуманитар ф. бүлеме. Казан : ТаРИХ, 2004. 111 б. ISBN 5-94113-099-6

13. Татар теле: үткәне һәм бүгенгесе. ТР ФА Г. Ибраһимов ис. Тел, әдәбият һәм сәнгать институты ; хезмәтне әзерләүчеләр: Р. Г. Галиуллин, Г. К. Һадиева. Казан, 2021. 72 б. ISBN 978-5-93091-369-9. EDN: ICNLSQ

14. Татар грамматикасы : өч томда / ТР ФА Г. Ибраһимов ис. Тел, әдәбият һәм сәнгать институты ; проект җит. М. З. Зәкиев ; ред. Ф. М. Хисамова. Тулыландырылган 2 нче басма. Казан, 2015-2017. Т. 1. 2015. 512 б. ISBN 978-5-93091-192-3. EDN: YUXEXM

15. Galieva A.M., Aminova A.A. [Metaphorical images of time in the poetic works of Su-leyman]. *Filologijaikul'tura*, 2020, no. 2 (60), pp. 18-26. (in Russ.). DOI: 10.26907/2074-0239-2020-60-2-18-26. EDN: HJKHAX

16. Can F., Patton J.M. Change of Writing Style with Time. Computers and the Humanities, 2004, vol. 38, pp. 61-82. DOI: 10.1023/B:CHUM.0000009225.28847.77

17. Takçı H., Ekinci E. Character Level Authorship Attribution for Turkish Text Documents. The Online Journal of Science and Technology, 2012, vol. 2, Issue 3, pp. 12-16.

18. Yülüce İ., Dalkılıç F. Author Identification with Machine Learning Algorithms. International Journal of Multidisciplinary Studies and Innovative Technologies, 2022, vol. 6, issue 1, pp. 45-50. DOI: 10.36287/ijmsit.6.1.45

19. Kocagül Yüzer H. Authorship Attribution in Turkish Texts. Ankara, Artsürem, 2022, 220 p. ISBN 978-605-72285-0-5

20. Galieva A.M., Aminova A.A. [Metaphorical images of time in the poetic works of Suleyman]. *Filologija i kul'tura*, 2020, no. 2 (60), pp. 18-26. (in Russ.). DOI: 10.26907/2074-0239-2020-60-2-18-26. EDN: HJKHAX

21. Stefan Evert [et al.]. Understanding and explaining Delta measures for authorship attribution.

Digital Scholarship in the Humanities, 2017, vol. 32, suppl. 2, pp. ii4-ii16. DOI: 10.1093/llc/fqx023

22. Burrows J. “Delta”: a Measure of Stylistic Difference and a Guide to Likely Authorship. Literary and Linguistic Computing, 2002, vol. 17, issue 3, pp. 267-287. DOI: 10.1093/llc/17.3.267

23. Stefan Evert [et al.]. Understanding and explaining Delta measures for authorship attribution. Digital Scholarship in the Humanities, 2017, vol. 32, suppl. 2, pp. ii4-ii16. DOI: 10.1093/llc/fqx023

24. *Rybicki J.*, *Eder M.* Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words? Literary and Linguistic Computing, 2010, vol. 26, issue 3, pp. 315-321. DOI: 10.1093/llc/fqr031

25. Velikanova N.P., Orehov B.V. [Digital textual criticism: text attribution using the example of M. A. Sholokhov’s novel “Quiet Don”]. *Mir Sholohova*, 2019, no. 1 (11), pp. 70-82. (in Russ.). EDN: HQDAHD

26. Alieva O.V. [Testing Burrows’ Delta on ancient Greek authors]. *Schole*, 2022, vol. 16, no. 2, pp. 693-705 (in Russ.). DOI: 10.25205/1995-4328-2022-16-2-693-705. EDN: SIQWWO

27. Eder M., Rybicki J., Kestemont M. Stylometry with R: A Package for Computational Text Analysis. The R Journal, 2016, vol. 8, no. 1, pp. 107-121. DOI: 10.32614/RJ-2016-007

28. Сүнчәләй С. Әсәрләр һәм хатлар. ТР ФА Г. Ибраһимов ис. Тел, әдәбият һәм сәнгать институты. Казан : Татарстан китап нәшрияты, 2005, 367 б. ISBN 5-298-04071-3

29. Stefan Evert [et al.]. Understanding and explaining Delta measures for authorship attribution. Digital Scholarship in the Humanities, 2017, vol. 32, suppl. 2, pp. ii4-ii16. DOI: 10.1093/llc/fqx023

M. A. Komyshev, Student
V. A. Baranov, Doctor of Philology, Professor
Kalashnikov Izhevsk State Technical University, Izhevsk, Russia

STYLOMETRIC ANALYSIS OF WORKS BYĞABDULLA TUQAYAND SĂĞİT SÜNÇÄLÄY: WHO IS THE AUTHOR OF THE POEM “SÖY ĞÖMERNE…”?

This article is devoted to determining the authorship of the Tatar quatrain “Söy ğömerne...” using stylometric methods. Ğabdulla Tuqay, to whom this text had been traditionally attributed, and Săğit Sünçäläy, evidence of whose authorship was discovered by literary scholars at the turn of the century, are considered possible authors. The paper tests several distance-based stylometric methods.

The material of the study was a corpus of digitized original poems by Ğabdulla Tuqay and Săğit Sünçäläy in the modern literary variety of the Tatar language. The final sample for analysis included 10 texts by Tuqay and 8 texts by Sünçäläy. The distance between texts was measured using Burrows’ Delta, Cosine Delta and cosine distance. Using the package Stylo, calculations were performed and dendrograms were constructed from them.

The Cosine Delta measure with extraction of character n-grams (with n = 4) proved to be the most effective. This is explained, firstly, by the agglutinative structure of the Tatar language, since among the most frequent n-grams extracted, elements corresponding to inflectional affixes were found. Secondly, the use of character n-grams allows to increase the size of the analyzed data. In some cases, good attribution results have also been achieved using Cosine Delta based on wordforms and cosine distance based on character n-grams.

Despite the impossibility to draw a definite conclusion due to the extremely small size of the texts analyzed, the comparison of the results of applying different methods shows that the most likely author of the quatrain in question is Săğit Sünçäläy.

Keywords: authorship; stylometry; Burrows’ Delta; Ğabdulla Tuqay; Săğit Sünçäläy; “Söy ğömerne…”; Tatar language.

Получена:16.04.2024
ГРНТИ 16.31.27

Образец цитирования

Комышев М. А., Баранов В. А. Стилометрический анализ произведений Габдуллы Тукая и Сагита Сунчелея: кто автор стихотворения «Сөй гомерне…»? // Социально-экономическое управление: теория и практика. 2024. Т. 20, № 2. С. 107–117. DOI: 10.22213/2618-9763-2024-2-107-117.

For Citation

Komyshev M.A., Baranov V.A. [Stylometric analysis of works by Ğabdulla Tuqay and Säğit Sünçäläy: who is the author of the poem “Söy Ğömerne…”?]. *Social’no-ekonomiceskoe upravlenie: teoria i praktika*, 2024, vol. 20, no. 2, pp. 107-117 (in Russ.). DOI: 10.22213/2618-9763-2024-2-107-117.