МАТЕМАТИЧЕСКАЯ МОДЕЛЬ И ЕЕ РЕАЛИЗАЦИЯ С ПРИМЕНЕНИЕМ КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ ДЛЯ АНАЛИЗА АВТОРСКОГО СТИЛЯ В КОНТЕКСТЕ ДИАЛОГА КУЛЬТУР*

С.Н. Дворяткина[1], С.А. Розанова[2]

[1] Елецкий государственный университет им. И.А. Бунина
ул. Коммунаров, 28, Елец, Россия, 399770
[2] Московский технологический университет (МИРЭА)
проспект Вернадского, 78, Москва, Россия, 119454

Статья посвящена актуальной проблеме диалога гуманитарной и естественно-научной культур в процессе решения важнейших междисциплинарных вопросов, связанных с изучением изменчивости авторских стилей под влиянием социокультурных условий. Предложена математическая модель анализа и сравнения стилей текстовых произведений. Раскрыта сущность технологии применения марковских цепей для анализа пар букв в их естественных последовательностях в тексте с целью установления устойчивости авторского стиля. На языке программирования C# разработан временной стилевой анализатор, обеспечивающий полный цикл проведения анализа стилей текстов. Программа позволяет прогнозировать изменение авторского стиля для любого временного периода; идентифицировать авторский стиль произведений путем введения параметра времени, что дает возможность сопоставить результаты статической таксономии (тексты берутся без учета временного параметра) и в динамике (с учетом времени создания).

Ключевые слова: интеграция естественнонаучного и гуманитарного знания, сравнение стилей, статистические методы идентификации текста, информационные технологии

В эпоху становления новой коммуникативной парадигмы современной постнеклассической науки, ориентированной на междисциплинарный подход и межкультурный диалог, требуется взвешенно и конструктивно подойти к решению задач, стоящих перед Россией и мировым сообществом в XXI веке — интеграции науки и культуры, технического и гуманитарного знания, совмещения современных достижений технического прогресса с культурными ценностями. Сегодня мир должен опираться на синергию знаний, диалог разных культур. Это направление исследования представляется нам наиболее актуальным.

Проблемой диалога занимались еще древнегреческие философы — Сократ, Платон, Аристотель. В философии второй половины XX века резко возросло внимание к проблемам диалога как основы творческого мышления. Ключевым понятием работы В.С. Библера «От наукоучения — к логике культуры. Два философских введения в двадцать первый век» является понятие диалога, которое автор связывает с мышлением, познанием. Попытку преодоления разрыва между гуманитарным и естественнонаучным знанием осуществил Чарльз Сноу. Ученый-физик и одновременно писатель впервые сформулировал проблему «двух

* Работа выполнена при поддержке РНФ, проект № 16-18-10304.

культур». Ряд российских ученых-педагогов в своих исследованиях, в той или иной степени, обращались к этой проблеме, внося в ее решение свой вклад. Так, например, в монографии [1] показана необходимость введения в учебный процесс по математике профессионально-прикладной и гуманитарной составляющих для гармонического развития личности и достижения мотивационного эффекта. На этой основе разработана концепция формирования математической культуры студентов и показаны пути ее реализации.

В исследовании [2] предложено следующее видение диалога естественнонаучной и гуманитарной культур и пути ее осуществления. Диалог естественнонаучной и гуманитарной культур рассматривается как их сближение, взаимопроникновение, взаимодействие и взаимообогащение. Очевидно, сочетание разных способов познания действительности — рационального естественнонаучного и иррационального гуманитарного, позволит решить проблему сохранения национальной идентичности, самобытных традиций, языка, уклада и духовно-нравственных ценностей русского народа, основу культурного многообразия в эпоху глобализации. В данной работе мы будем исходить из этого понимания диалога культур.

Проблема анализа и сравнения стилей текстовых произведений давно уже носит междисциплинарный характер благодаря эффективному привлечению математических методов. Математические методы позволяют получить не только количественные, но и качественные выводы в филологических исследованиях. В литературоведческой практике проверка текстов на близость стилей необходима для установления в спорных случаях подлинного авторства литературных произведений, особенно удаленных временем. В качестве примеров можно привести споры об авторстве некоторых произведений Шекспира, отдельных анонимных и псевдонимных публицистических статей, приписываемых Ф.М. Достоевскому, стихотворных текстов М.Ю. Лермонтова, прозаических произведений М.Е. Салтыкова-Щедрина, М.А. Шолохова и т.д.

Теоретическим основанием для использования математических методов исследования письменной речи является статистическая модель порождения речевого высказывания. Благодаря прочной фиксации навыков письма и образования в коре головного мозга систем динамического стереотипа, труд, затрачиваемый в процессе создания текста, уменьшается, и появляется возможность писать автоматизировано. Этот фактор позволяет доподлинно определять авторство.

Вопросами проверки текстов произведений на близость стилей с применением формально-количественных методов в российской и зарубежной науке занимались А.А. Марков, М.А. Марусенко, Н.А. Морозов, В.П. Фоменко, Т.Г. Фоменко, В.Фукс, Д.В. Хмелев, Г. Хетсо, О.Г. Шевелев и др. [3—8]. С развитием информационных технологий анализ текстов приобрел огромный научный интерес. В 1978 году математик Г. Хетсо [7] предложил методику установления авторства, основанную на анализе текста с автоматизированным получением частотных словарей и статистических данных. К сожалению, в разработке была допущена ошибка, заключающаяся в использовании только одного параметра — средней длины предложения.

Одним из последних исследований, основанных на автоматической обработке текста, является диссертационная работа О.В. Шевелева [9], в которой был предложен новый подход для сравнения стилей текстов, базирующийся на двустороннем критерии Фишера и χ^2-критерии Пирсона по частотным признакам, совокупности признаков и их распределению. Автором разработан программный комплекс «СтилеАнализатор», обеспечивающий полный цикл проведения анализа стилей текстов.

Несмотря на множество работ по проверке на близость стилей текстов, все же остается еще ряд мало исследованных областей. Например, ни в одной из существующих работ практически не поднимался вопрос о применении математических методов для прогноза изменения (сохранения) стиля автора, тексты которого были созданы в разные периоды жизни писателя под воздействием объективных социокультурных факторов.

Сравнение конкретных текстов возможно на основе совокупности признаков, отражающих существенные свойства авторского стиля. К идентифицирующим признакам можно отнести: статистические характеристики (частотность слов, букв, их сочетаний, количественное использование определенных частей речи, синтаксических конструкций и т.д.). В науке установлено также, что вероятность появления сочетания пар символов различной природы в тексте отдельных авторов подчиняются некоторым устойчивым закономерностям. При этом основная проблема формальных методов анализа авторских стилей состоит в выборе необходимых компонентов. Характеризующие параметры, по замечанию А.А. Маркова [3], должны удовлетворять определенным требованиям, таким, как статистическая устойчивость, массовость, различающая способность и, следовательно, могут быть формализованы только с помощью количественного анализа текстовых единиц с применением вероятностно-статистических методов.

Нами предлагается технология применения марковских цепей для анализа пар букв в их естественных последовательностях в тексте с целью установления устойчивости авторского стиля, сущность которой состоит в следующем. Пусть имеются достаточно длинные фрагменты (не более 100 000 символов) прозаических произведений одного автора на русском языке, написанные в разные периоды жизни. Например, произведения Ивана Алексеевича Бунина «Деревня», изданное в 1910 г. в России, и роман «Жизнь Арсеньева» 1927 г., созданный после эмиграции во Францию. По произведениям раннего периода (выбирается одно контрольное произведение) вычисляется матрица переходных вероятностей встречаемости пар букв, которая служит оценкой матрицы вероятностей перехода из буквы в букву для экспериментального произведения позднего периода. Если вычисленная оценка вероятности высока, то стиль автора под воздействием внешних факторов не изменился, и наоборот. Такой метод оказывается достаточно точным для естественно-языковых текстов. Данное исследование проводим формальными методами анализа текста с применением аппарата марковских цепей.

Рассмотрим подробнее математическую модель для определения авторского стиля.

1. Предположим, что вероятности перехода p_{ij} из одной буквы в другую являются реализацией цепи Маркова для раннего произведения с переходной матри-

цей $\mathcal{P}$. Данные вероятности вычисляются по формуле условных вероятностей: $p_{ij} = \frac{p(ij)}{p(j)}$, где $p(ij)$ — вероятность встречаемости пар букв i и j, а $p(j)$ — вероятность встречаемости буквы j в тексте.

2. Полученную матрицу переходных вероятностей возводим в степень m, т.е. находим $\mathcal{P}^m$, где m — это временной период с года написания раннего произведения до позднего. Построенная матрица $\mathcal{P}^m$ является прогнозируемой теоретической матрицей переходных вероятностей.

3. Далее строим эмпирическую матрицу переходных вероятностей $\mathcal{P}^1$ для позднего произведения, согласно п. 1.

4. Осуществляем статистическую проверку теоретической матрицы $\mathcal{P}^m$ с эмпирической матрицей $\mathcal{P}^1$ по χ^2-критерию Пирсона.

Перейдем к сравнению построенных нами ранее матриц переходных вероятностей $\mathcal{P}^m$ (матрица переходов для раннего произведения, возведенная в степень m) и $\mathcal{P}^1$ (матрица переходов для позднего произведения).

4.1. Формулируем нулевую и альтернативные гипотезы: H_0 — распределение признака по теоретической матрице совпадает с распределением признака по эмпирической матрице; H_1 — распределение признака по теоретической матрице значимо отличается от распределения признака по эмпирической матрице.

4.2. Задаем уровень значимости $\alpha = 0{,}05$.

4.3. Находим эмпирическое значение критерия по формуле: $\chi^2_{\text{эм}} = \sum \frac{(np - np')^2}{np'}$, где np — эмпирическая частота, np' — теоретическая частота.

4.4. Определяем критическое значение статистики Пирсона для $\alpha = 0{,}05$ и числа степеней свободы, равного $k = 33^2 - 1 = 1088$. Имеем $\chi^2_{\text{кр}}(0{,}05; 1088) = 1012{,}425$.

4.5. Делаем статистические выводы. Если $\chi^2_{\text{эм}} < \chi^2_{\text{кр}}(\alpha; k)$, то нет оснований отвергнуть нулевую гипотезу H_0. Если эмпирическое значение критерия $\chi^2_{\text{эм}}$ попало в критическую область $\chi^2_{\text{эм}} \geq \chi^2_{\text{кр}}(\alpha; k)$, то нулевую гипотезу H_0 отвергают.

Чтобы получить более точные результаты, перейдем к программе реализации модели на языке C# для сравнения авторского стиля путем разработки временного стилевого анализатора. Предложенная программа позволяет прогнозировать изменение авторского стиля для любого временного периода; идентифицировать авторский стиль произведений путем введения параметра времени, что дает возможность сопоставить результаты как статической таксономии (тексты берутся без учета временного параметра), так и в динамике (с учетом времени создания).

Общий вид программы «Временной стилевой анализатор» содержит два окна «Текст № 1» и «Текст № 2», которые служат для ввода сравниваемых текстов. Разберем на примере первого фрагмента работу стилевого анализатора.

1. В поле «Текст № 1» помещаем фрагмент произведения «Деревня» размером 2765 символов и нажимаем кнопку «Анализировать № 1» (рис. 1). Следует уточнить, что для корректной работы программы вводимый в поле текст должен иметь размер не более 100 000 символов.

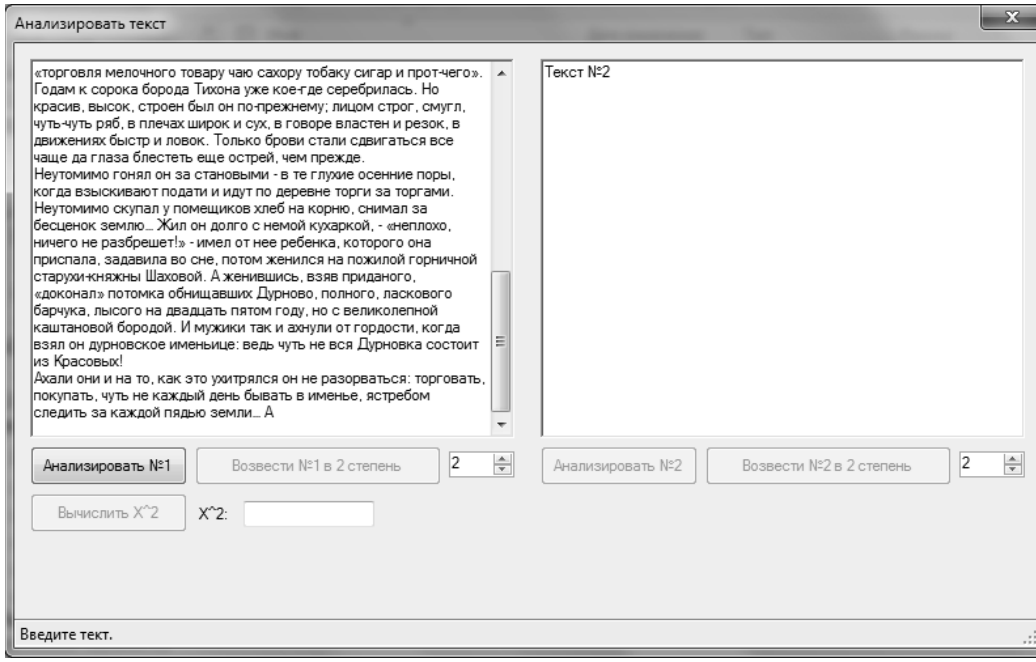

Рис. 1. Шаг № 1 алгоритма программы

2. После чего получаем матрицу переходных вероятностей, которую необходимо возвести в $m = 17$ степень. По умолчанию в поле для степени прописано значение 2, меняем его на 17 и нажимаем кнопку «Возвести № 1 в 17 степень». На экране отобразится окно с матрицей размером $n \times n$, где $n = 33$, т. е. количество букв русского алфавита (рис. 2).

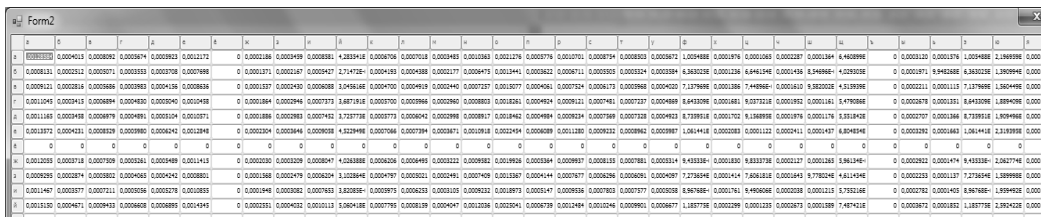

Рис. 2. Шаг № 2 алгоритма программы

3. Далее необходимо найти матрицу переходных вероятностей для отрывка из романа «Жизнь Арсеньева». Введем в поле «Текст № 2» часть текста размером

2765 символов и нажмем кнопку «Анализировать № 2» (рис. 3). Следует отметить, что количество символов в произведениях должно быть одинаковым для корректного сравнения, так как χ^2-критерий Пирсона работает при равном объеме выборки.

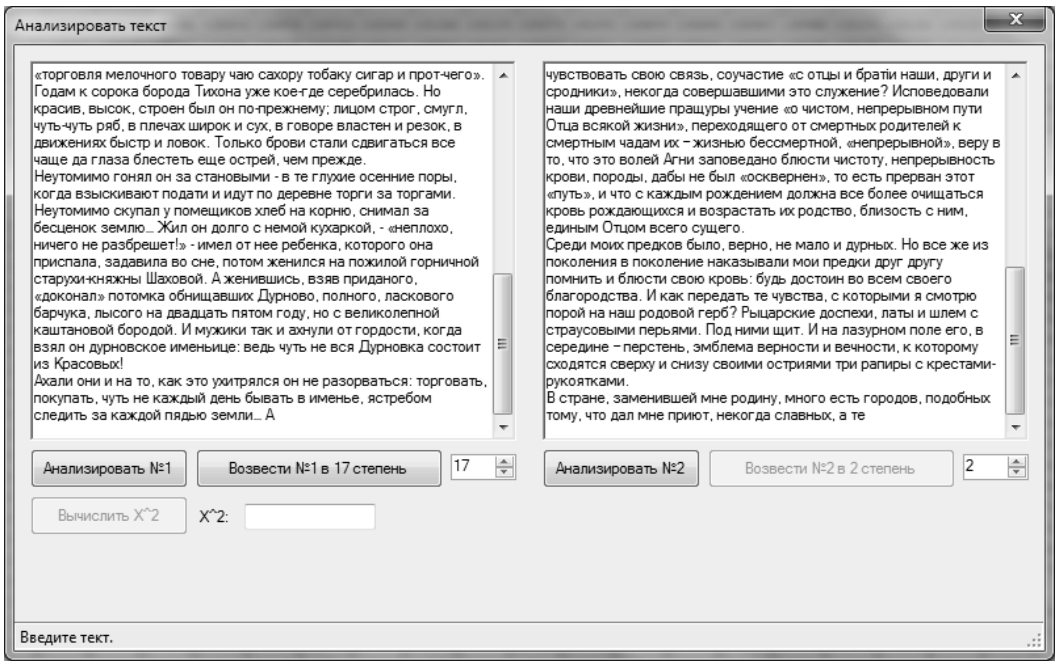

Рис. 3. Шаг № 3 алгоритма программы

4. Для процедуры количественного сравнения стилей данных произведений достаточно нажать на кнопку «Вычислить χ^2», после чего на экране отобразится результат вычислений (рис. 4). Сравнивая значение, вычисленное программой и отображенное в окне с именем «Результат», $\chi^2_{эмп}$ с ранее полученным критическим значением $\chi^2_{кр}$, можно заключить, что $\chi^2_{эмп} = 1569{,}801$ значимо больше $\chi^2_{кр}(0{,}05; 1088) = 1012{,}425$. Следовательно, отклоняем нулевую гипотезу и делаем вывод о том, что стилистические особенности И.А. Бунина изменились под действием социокультурной среды.

Сравнение «ранних» и «поздних» произведений позволяет утверждать, что под влиянием социокультурной среды русского зарубежья 20-х годов прошлого столетия произошли изменения стиля выдающегося писателя И.А. Бунина. Литературоведы теперь могут с уверенностью заявлять, что существуют объективные факторы, повлиявшие на субъективные воплощения авторских переживаний в слове, на стилевую ткань произведений. Если к этим наблюдениям прибавить количественные показатели, полученные в процессе статистического анализа, то выводы, касающиеся изменения авторского стиля, становятся более достоверными — математически подтвержденными [10; 11]. Таким образом, диалог естественнонаучных и гуманитарных знаний обогащает наше представление о русских художниках слова, об особенностях творческого процесса как сложного духовного явления и в конкретном случае оказывается весьма продуктивным.

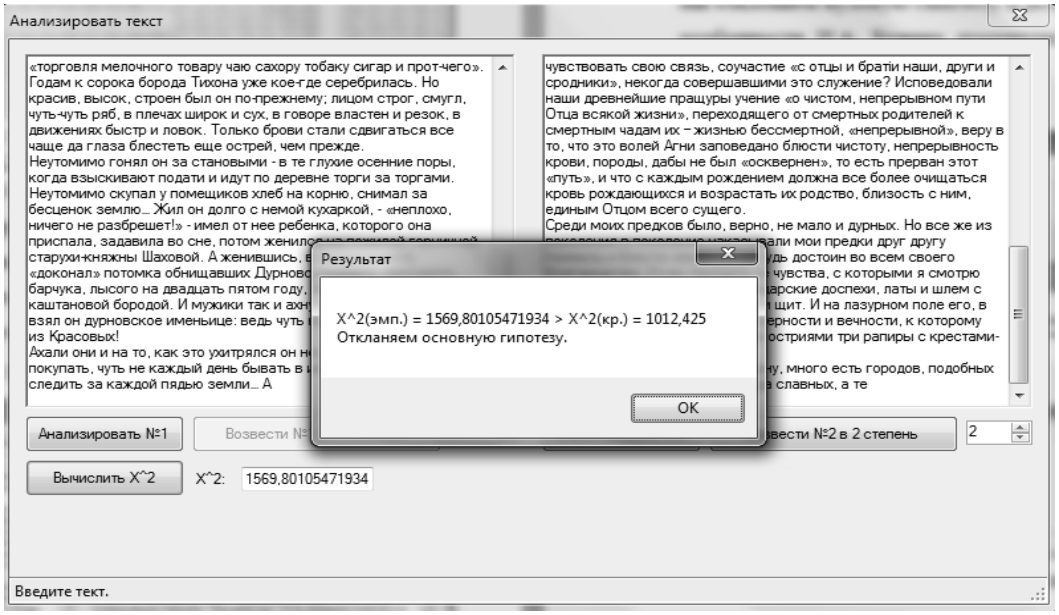

Рис. 4. Шаг № 4 алгоритма программы

В заключение следует отметить, что разработанный стилевой анализатор может использоваться не только в научной деятельности лингвистов, филологов, историков, культурологов, криминалистов для проверки текстов на стилистическую идентификацию, на установление авторства и стилистических особенностей языка литературных произведений различных жанров, созданных в разные временные периоды, но и в образовательной сфере как для гуманитарных, так и для естественнонаучных, инженерно-технических направлений подготовки и специальностей. Задачи, аналогичные приведенной в данной статье, целесообразно включать в интегративные материалы и курсы, на семинарские занятия и в проектную деятельность студентов по лингвистике, стилистике, лексикологии, литературе, прикладной математике и информационным технологиям. Их решение способствует повышению учебной и профессиональной мотивации студентов и, следовательно, повышению качества образования.

ЛИТЕРАТУРА

[1] *Розанова С.А.* Математическая культура студентов технических университетов: монография. М.: ФИЗМАТЛИТ, 2003. 176 с.

[2] *Дворяткина С.Н.* Развитие вероятностного стиля мышления в процессе обучения математике: теория и практика: монография. М.: ИНФРА-М, 2013. 272 с.

[3] *Марков А.А.* Пример статистического исследования над текстом «Евгения Онегина», иллюстрирующий связь испытаний в цепь // Известия Имп. Акад. Наук. Серия VI, Т. X, N3, 1913.

[4] *Морозов Н.А.* Лингвистические спектры. Средство для отличения плагиатов от истинных произведений того или иного известного автора: Стилеметрический этюд // Известия отд. русского языка и словесности Имп. Акад. Наук. 1915, Т. XX, Кн. 4.

[5] *Марков А.А.* Об одном применении статистического метода // Известия отд. русского языка и словесности Имп. Акад. Наук. 1916, Серия VI, Т.X.

[6] *Фукс В.* По всем правилам искусства: Точные методы в исследованиях литературы, музыки и изобразительного искусства // Искусство и ЭВМ / под ред. Р.Х. Зарипова. М.: Мир, 1975. С. 134—356.
[7] *Хетсо Г.* Проблема авторства в романе «Тихий дон» // Scando-slavica. 1978, Т. 24.
[8] *Марусенко Н.А.* Атрибуция анонимных и псевдонимных литературных произведений методами распознавания образов. Л.: Изд-во ЛГУ, 1990.
[9] *Шевелев О.В.* Разработка и исследование алгоритмов сравнение стилей текстовых произведений: дисс. ... канд. техн. наук. Томск, 2006. 176 с.
[10] *Дворяткина С.Н., Дякина А.Н., Мельникова Ю.В.* Аппарат цепей Маркова в анализе изменений авторского стиля под воздействием социокультурной среды: к постановке проблемы // Вестник Елецкого государственного университета им. И.А. Бунина. Вып. 34: Серия «Педагогика» (История и теория математического образования). Елец: ЕГУ им. И.А. Бунина, 2014. С. 159—164.
[11] *Dvoryatkina S.N., Dyakina A.A.* On Variability of Authors' Style under the Influence of the Socio-Cultural Environment in the Context of Dialogue of Natural Scientific and Humanitarian Cultures // Mediterranean Journal of Social Sciences MCSER Publishing, Rome-Italy. Vol 6, No 5 S4 October 2015. Special Issue. P. 167—171.

MATHEMATICAL MODEL AND ITS REALIZATION WITH THE USE COMPUTER TECHNOLOGIES FOR THE ANALYSIS OF THE AUTHOR'S STYLE IN THE CONTEXT OF INTERCULTURAL DIALOGUE

S.N. Dvoryadkin[1], S.A. Rozanova[2]

[1] Yelets State Bunin University
Kommunarov str., 28, Yelets, Russia, 399770
[2] Moscow Technological University
prospect Vernadskogo, 78, Moscow, Russia, 119454

The article is devoted to an actual problem of natural-science and humanitarian cultures dialogue in the process of solving the most important cross-cutting issues associated with the study of writing styles variability under the influence of sociocultural conditions. There is offered a mathematical model of analyzing and comparing the text works styles. The essence of technology of Markov chains application to analyzing of the letter pairs in their natural sequences in the text in order to establish the stability of the author's style. Temporary style analyzer, which provides a full cycle analysis text styles, was developed in the C # programming language. The program let the opportunity to predict changes in the author's style for any time period; identify the author's style works by introducing a time parameter that allows comparing the results of the static taxonomy (the texts are taken without regard to the the time parameter) and in dynamics (including the creation time).

Key words: integration of science and human knowledge, comparison of styles, statistical methods for the identification and information technology

REFERENCES

[1] *Rozanova S.A. Matematicheskaya kul'tura studentov tekhnicheskikh universitetov. Monografiya* [Mathematical culture of students of technical universities. Monograph]. M.: FIZMATLIT, 2003. 176 p.

[2] *Dvorjatkina S.N. Razvitie verojatnostnogo stilja myshlenija v processe obuchenija matematike: teorija i praktika: Monografija* [Development of probabilistic thinking styles in the educational process of mathematics: Theory and Practice: Monograph]. M.: INFRA-M, 2013. 272 p.

[3] *Markov A.A. Primer statisticheskogo issledovaniya nad tekstom «Evgeniya Onegina», illyustriruyushchiy svyaz' ispytaniy v tsep'* [An Exmple of Statistical Research of the Text of "Evgeniy Onegin" Illustrating the Connection of the Tests in a Chain]. *Izvestiya otd. russkogo yazyka i slovesnosti* [News of the Imp. Acad. Of Sciences]. 1913. Series I, Vol. X, No. 3.

[4] *Morozov N.A. Lingvisticheskie spektry. Sredstvo dlya otlicheniya plagiatov ot istinnykh proizvedeniy togo ili inogo izvestnogo avtora: Stilemetricheskiy etyud* [Linguistic Spectrum. The Means for the Distinction of Authentic Works of This or That Famous Writer: Style and Metric Essay]. *Izvestiya otd. russkogo yazyka i slovesnosti Imp. Akad. Nauk* [News of the Dept. Of Russian Language and Literature of the Imp. Acad. of Science]. 1915. Vol. XX. Book 4.

[5] *Markov A.A. Ob odnom primenenii statisticheskogo metoda* [On One Application of the Statystical Method]. *Izvestiya otd. russkogo yazyka i slovesnosti Imp. Akad. Nauk* [News of the Dept. Of Russian Language and Literature of the Imp. Acad. of Science]. 1916. Series VI, Vol. X.

[6] *Fucks W. Po vsem pravilam iskusstva: tochnye metody v issledovaniyah literatury, muzyki i izobrazitelnogo iskusstva* [By All Rules of Art: Accurate Methods in the Literature, Music and Visual Art Researches]. *Iskusstvo i EV. Pod red. R.Kh. Zaripova* [In Art and Art and Computers, edited by Zaripova R.H.]. Moscow: Mir, 1975. Pp. 134—356.

[7] *Kjetsaa G. Problema avtorstva v romane "Tikhiy Don"* [The Problem of Authorship in the Novel "And Quiet Flows the Don"]. 1978. Scando-slavica. Vol. 24. Pp. 91—105.

[8] *Marusenko N.A. Atributsiya anonimnyh i psevdonimnyh literaturnyh proizvedeniy metodami raspoznavaniya obrazov* [Attribution of Anonimous and Pseudonimic Literature Works by Image Recognition Methods]. St. Petersburgh: LGU, 1990.

[9] *Shevelev O.V. Razrabotka i issledovanie algoritmov sravneniye stiley tekstovyh proizvedeniy* [Development and Study of Algorithms for Comparison of the Text Works Styles]. *Dissertatsiya kandidata tekhnicheskikh nauk* [Dissertation of the Candidate of Technical Science]. Tomsk, 2006.

[10] *Dvoryatkina S.N., Dyakina A.N., Melnikova Yu.V. (2014). Apparat tsepey Markova v analize izmeneniy avtorskogo stilya pod vozdeystviyem sotsiokulturnoy sredy: k postanovke problemy* [Markov's Chains Apparatus in Analysis of Author's Style Variability under the Influence of Socio-Cultural Environment: to the Problem Setting]. *Vestnik Eletskogo gosudarstvennogo universiteta im. I.A. Bunina. Vyp. 34: Seriya «Pedagogika» (Istoriya i teoriya matematicheskogo obrazovaniya)* [Bulletin of Yeletskiy State University Named after I.A. Bunin. Issue No. 34: "Pedagogika" Series (History and Theory of Mathematical Education)]. Yelets: I.A. Bunin EGU. Pp. 159—164.

[11] *Dvoryatkina S.N., Dyakina A.A.* On Variability of Authors' Style under the Influence of the Socio-Cultural Environment in the Context of Dialogue of Natural Scientific and Humanitarian Cultures // Mediterranean Journal of Social Sciences MCSER Publishing, Rome-Italy. Vol 6, No 5 S4 October 2015. Special Issue. Pp. 167—171.