УДК 655.5

М. А. Зильберглейт, доктор химических наук, профессор (БГТУ);
Ю. Ф. Шпаковский, кандидат филологических наук, доцент (БГТУ);
М. М. Невдах, кандидат технических наук (БГТУ)

РАЗРАБОТКА МОДЕЛИ КОНТРОЛЯ КАЧЕСТВА УЧЕБНЫХ МАТЕРИАЛОВ ПРИ ПОДГОТОВКЕ К ПЕЧАТИ

В статье приведены результаты исследования, посвященного автоматизированному контролю качества учебных материалов на допечатной стадии полиграфического производства. Для достижения цели в работе на первом этапе проведены эксперименты с использованием различных методик для получения объективных критериев относительно качества текстов. На втором этапе выделены и вычислены значения 49 параметров издательских оригиналов. Снижение признакового пространства осуществлялось методами многомерного статистического анализа. Для разработки решающего правила использовался дискриминантный анализ. Для автоматизированного контроля качества материалов создана программа Readability analysis.

The article presents the results of a study on automated quality control of teaching materials in the prepress stage of the printing industry. To achieve the goal in the first stage of the work carried out experiments using different methods to obtain objective criteria about the difficulty of texts. In the second stage have been identified and calculated values of 49 parameters of originals. Reducing the feature space is carried out by multivariate statistical analysis. To develop a decision rule used discriminant analysis. For automated quality control of materials created program Readability analysis.

Введение. В последнее время постоянно повышается уровень управления различными этапами полиграфического производства, однако допечатные процессы ввиду своей сложности и разнообразия не позволяют осуществлять эффективное решение задач планирования, управления, моделирования и автоматизации полиграфического процесса. В связи с этим особую актуальность приобретает проблема создания адекватных экономических, технических, математических и других формализованных моделей на допечатной стадии полиграфического производства.

Как известно, задачей допечатных процессов является обеспечение надлежащего качества подготовки издательской продукции к полиграфическому воспроизведению. С этой целью в Республике Беларусь утвержден ряд нормативных документов. В частности, приняты различные стандарты (СТБ 1339-2002 «Учебники и учебные пособия для системы высшего и среднего специального образования. Общие технические условия», СТБ 1021-2004 «Издания учебные для общего среднего образования. Общие технические условия», ГОСТ 7.89–2005 «Оригиналы текстовые авторские и издательские. Общие требования», СТБ 7.204-2006 «Издания книжные. Общие технические условия», СТБ 7.206-2006 «Издания книжные и журнальные для детей. Общие технические условия» и др.), однако основное внимание в них уделено общим техническим требованиям.

В связи с этим разработка модели контроля качества учебных материалов на допечатной стадии полиграфического производства является актуальной проблемой.

Основная часть. Следует отметить, что решение некоторых задач (атрибуция текста, оценка близости и однородности стилей, их классификация, читабельность) отражены в работах отечественных и зарубежных исследователей: Гринбаума О. Н., Мартыненко Г. Я., Поддубного В. В., Шевелева О. Г., Мацковского М. С., Пиотровского Р. Г., Хмелева Д. В., Флеша Р., Чолл Дж. и др. [1–10]. Кроме того, для решения конкретных прикладных задач разработано программное обеспечение: «ЛингвоАнализатор», «СМАЛТ», «ЛинДа», PolyAnalyst, DICTUM, «ВААЛ» и др.

Однако специальных исследований, нацеленных на разработку модели контроля качества издательской продукции, не предпринималось. Не сформулированы и основные подходы к разработке методики ее использования.

В связи с этим была определена *цель исследования* — разработка модели контроля качества учебного материала на допечатной стадии полиграфического производства.

Для реализации указанной цели были поставлены и решены следующие задачи:

— провести классификацию, анализ и отбор методов проверки качества печатного материала на допечатной стадии полиграфического производства, установить объективные показатели качества исследуемых объектов выборки;

— определить и измерить количественные характеристики исходного материала и экспе-

риментальным путем выявить их связь с качеством; снизить размерность признакового пространства методами многомерного статистического анализа и выявить наиболее информативные показатели;

— на основе экспертных данных и информативных признаков с помощью дискриминантного анализа разработать модель процесса в виде решающего правила для контроля качества учебных материалов;

— разработать специализированное программное обеспечение для оценки качества авторского оригинала и управления качеством издательских проектов.

Общая схема исследования представлена на рис. 1.

Из приведенной схемы следует, что конечной целью диссертационной работы является получение решающего правила в виде набора дискриминантных (классифицирующих) функций для контроля качества учебных материалов и разработка специализированного программного обеспечения для внедрения в существующий технологический процесс издательских предприятий.

На *первом этапе* исследования были проведены эксперименты с использованием различных методик. В экспертизе участвовало 75 реципиентов, что позволило с вероятностью 99% получить относительную ошибку в долях среднеквадратичного отклонения, равную 0,3. Экспериментальными материалами послужили издательские оригиналы для вузов по философии (первая выборка) и экономической теории (вторая выборка). Каждая выборка содержала 24 отрывка издательского оригинала.

В работе использовались наиболее надежные методы: методика дополнения, метод балльных оценок. Впервые для оценки качества учебного материала для вузов использовался метод парных сравнений. Для выявления связи между мнениями экспертов в последнем методе рассчитывался коэффициент конкордации. Оценка его значимости осуществлялась на основе χ^2-критерия Пирсона.

Обработка и анализ результатов экспериментов позволили выявить информацию относительно качества материалов. На основании полученных данных найдены пять объективных показателей качества: процент правильно заполненных пропусков (Y_1); относительное время работы с текстом (Y_2) — с использованием методики дополнения; средняя оценка качества текста (Y_3); относительное время работы с текстом (Y_4) — с использованием балльных оценок; ранг текста (Y_5).

Для каждого показателя была найдена середина диапазона всех полученных значений, в соответствии с которой производилось разбиение объектов на два класса, которые мы условно назвали «материал требует доработки — материал не требует доработки». В итоге было получено разбиение текстов на группы по выделенным пяти показателям качества.

Второй этап исследования посвящен изучению информационных характеристик исследуемых объектов и выявлению объективных диагностических показателей, которые в наибольшей степени влияют на качество учебных материалов.

С этой целью были выделены и вычислены значения 49 параметров учебных текстов (длина текста в абзацах, длина текста в словах, длина текста в буквах, средняя длина абзаца в фразах, средняя длина абзаца в словах, средняя длина абзаца в буквах, средняя длина абзаца в печатных знаках, средняя длина предложения в фразах и др.).

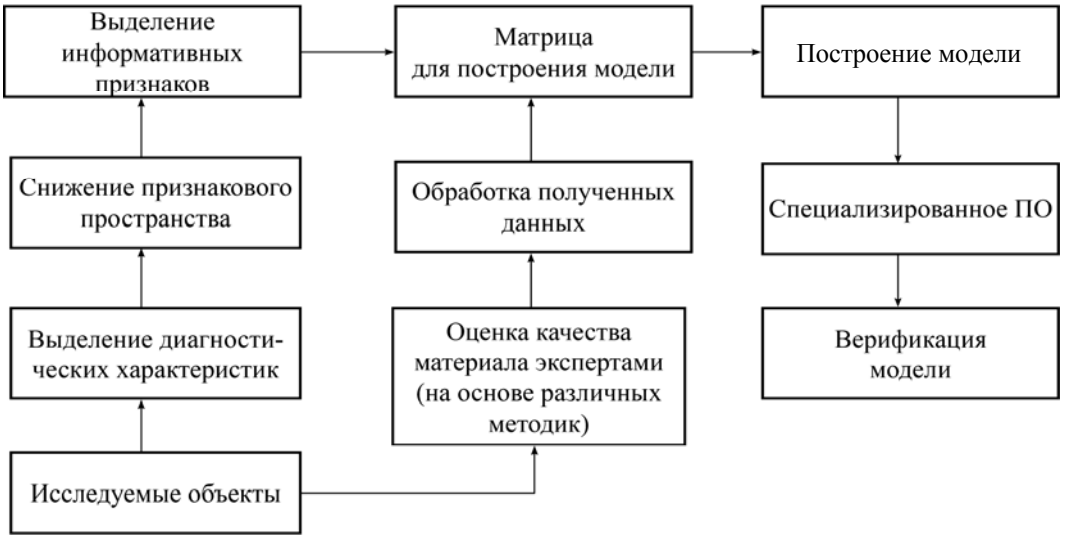

Рис. 1. Общая схема исследования

Очевидно, что использование большого количества показателей является неэффективным по ряду причин: а) сильная взаимосвязанность признаков; б) неинформативность признаков, мало меняющихся при переходе от одного объекта к другому (малая «вариабельность» признаков); в) возможность агрегирования по некоторым признакам. Для снижения признакового пространства использовались методы многомерного статистического анализа (кластерный и факторный анализ, метод корреляционных плеяд и вроцлавской таксономии, многомерное шкалирование).

В кластерном анализе для нахождения расстояния между объектами использовались следующие меры сходства: Евклидово расстояние, квадрат расстояния Евклида, косинус угла, коэффициент корреляции Пирсона, неравенство Чебышева, расстояние Минковского. Для кластеризации использовались следующие способы: метод простого среднего, метод группового среднего, метод ближнего соседа, метод дальнего соседа, невзвешенный центроидный метод, взвешенный центроидный метод, метод Варда.

В результате кластеризации выделенных характеристик печатного материала была получена информация о формировании кластеров: порядок объединения кластеров, расстояние между ними, а также принадлежность характеристик объекта к тому или иному кластеру.

Кластерный анализ позволил выделить шесть групп для первой выборки и девять групп — для второй.

Для снижения признакового пространства в работе использовались следующие варианты факторного анализа: метод главных факторов, центроидный метод и метод главных компонент.

При проведении факторного анализа было установлено, что первые три фактора объясняют около 74% разброса дисперсии для первой выборки, около 64% — для второй. Для решения вопроса, какие из факторов следует оставить для дальнейшей обработки, использовались критерий Кайзера и критерий «каменистой осыпи» Р. Кэтелла.

С целью получения более простой структуры, которой соответствует большое значение нагрузки каждой переменной только по одному фактору, в работе использовались ортогональные методы вращения: варимакс, квартимакс и эквимакс.

Изучение результатов с использованием всех методов факторного анализа и методов вращения позволило выявить, как признаки распределились между факторами. Анализ показал, что факторы по всем методам вращения для двух выборок практически идентичны.

Результаты, полученные методами факторного анализа, позволили выделить шесть групп для первой выборки и восемь групп — для второй.

Наряду с методами факторного анализа для снижения признакового пространства использовался метод корреляционных плеяд. Упорядочение производилось на основании принципа максимального корреляционного пути. Для удобства построения графа были составлены упорядоченные корреляционные матрицы. На основании упорядочения всех признаков были построены графы, которые представляют собой кратчайший незамкнутый путь. После выбора порогового значения коэффициента корреляции исходный граф распался на пять подграфов (групп близких параметров) для первой выборки и шесть подграфов — для второй выборки.

В методе вроцлавской таксономии (методе дендритов) точки многомерного пространства проецировались на плоскость, чем достигалось нелинейное упорядочение изучаемых элементов. Из дендрита, построенного на единицах разбиваемого множества, удалялось (n – 1) самых длинных связей. Тем самым получалось разбиение дендрита на n частей, которое характеризовалось минимальной суммой образующих их отрезков, а полученные подмножества включали элементы с близкими значениями признаков.

С целью построения дендрита вычислены матрицы расстояний (на основе расстояния Евклида) между изучаемыми характеристиками. Данный метод позволил выделить семь групп для первой выборки и пять групп — для второй.

При многомерном шкалировании матрица различий между объектами представлялась в пространстве относительно небольшого числа измерений с наименьшим возможным искажением геометрической структуры исходных данных. Качество метода характеризовалось величиной коэффициентов стресса и R^2.

Наилучшей моделью для первой выборки (stress = 0,210, R^2 = 0,856) стала модель, полученная с использованием меры сходства, основанной на неравенстве Чебышева; для второй выборки (stress = 0,230; R^2 = 0,763) — модель, полученная с использованием квадрата расстояния Евклида. На их основе было получено семь групп признаков для первой выборки и шесть групп — для второй.

Сравнение результатов для учебных текстов по философии и экономической теории, полученных с помощью разных методов многомерного статистического анализа, позволило сделать следующий вывод: во многих случаях совпадают не только отдельные признаки в группах, но и сами группы.

Таким образом, впервые в области контроля качества учебных материалов на подготовительной стадии полиграфического процесса методами многомерного статистического анализа установлены диагностические признаки, т. е. те показатели исследуемых объектов, которые в наибольшей степени влияют на качество. Ими оказались следующие признаки: длина слов и предложений (кластерный анализ); сложность предложений, число предикативных ядер, длина фразы (факторный анализ); разнообразие словаря, длина абзаца, слов и предложений, средняя длина фразы и предложения (метод корреляционных плеяд); длина слов и предложений (методы вроцлавской таксономии и многомерного шкалирования).

Для дальнейшего изучения характеристик учебных текстов важнейшей задачей является выделение наиболее информативного признака из каждой полученной группы. В данной работе для оценки информативности признаков использовалась мера С. Кульбака.

На основе данной меры были вычислены информационные меры каждого из 49 исследуемых признаков, а затем отобраны те из них, которые обладают наибольшей информативностью среди признаков своей группы. В результате число признаков было сокращено до возможного минимума.

На третьем этапе на основе диагностических признаков и экспертных данных был проведен дискриминантный анализ, который позволил разработать решающее правило для автоматической проверки качества материалов на подготовительной стадии полиграфического процесса. Точность классификации объектов первой выборки составила 91,7% (табл. 1), второй — 83,3% (табл. 2).

Таблица 1

Результаты разбиения объектов первой выборки на классы

№ объекта	Разбиение на классы в результате		Результат	№ объекта	Разбиение на классы в результате		Результат
	эксперимента	дискриминантного анализа			эксперимента	дискриминантного анализа	
1	1	1	Верно	13	1	1	Верно
2	1	1	Верно	*14	0	1	**Неверно**
3	1	1	Верно	15	1	1	Верно
4	0	0	Верно	16	0	0	Верно
5	0	0	Верно	17	0	0	Верно
6	1	1	Верно	18	0	0	Верно
7	1	1	Верно	19	0	0	Верно
8	0	0	Верно	*20	1	0	**Неверно**
9	1	1	Верно	21	1	1	Верно
10	0	0	Верно	22	0	0	Верно
11	0	0	Верно	23	0	0	Верно
12	1	1	Верно	24	1	1	Верно

Таблица 2

Результаты разбиения объектов второй выборки на классы

№ объекта	Разбиение на классы в результате		Результат	№ объекта	Разбиение на классы в результате		Результат
	эксперимента	дискриминантного анализа			эксперимента	дискриминантного анализа	
*1	1	0	**Неверно**	13	0	0	Верно
2	0	0	Верно	14	0	0	Верно
3	0	0	Верно	*15	1	0	**Неверно**
4	0	0	Верно	16	1	1	Верно
5	0	0	Верно	17	1	1	Верно
6	0	0	Верно	18	1	1	Верно
7	0	0	Верно	19	1	1	Верно
*8	0	1	**Неверно**	20	1	1	Верно
9	0	0	Верно	21	1	1	Верно
*10	0	1	**Неверно**	22	1	1	Верно
11	1	1	Верно	23	1	1	Верно
12	0	0	Верно	24	0	0	Верно

С помощью дискриминантного анализа для объектов первой выборки для дальнейшей программной реализации получены следующие дискриминантные функции:

$$F_1 = -53{,}062 - 0{,}015X_3 + 0{,}831X_9 - 15{,}106X_{24};$$

$$F_2 = -42{,}720 - 0{,}011X_3 + 0{,}554X_9 - 8{,}663X_{24},$$

где X_3 — длина текста в буквах; X_9 — средняя длина предложения в словах; X_{24} — средняя длина слов в печатных знаках.

Для объектов второй выборки получены следующие дискриминантные функции:

$$F_1 = -123{,}728 - 0{,}165X_5 + 0{,}268X_{10} - 3{,}100X_{39};$$

$$F_2 = -104{,}608 - 0{,}100X_5 + 0{,}2229X_{10} - 2{,}830X_{39}.$$

где X_5 — средняя длина абзаца в словах; X_{10} — средняя длина предложения в слогах; X_{39} — процент неповторяющихся слов.

Очевидно, что в зависимости от происхождения изучаемой выборки факторы, влияющие на качество, различны. Дискриминантный анализ подтвердил этот факт и позволил выявить некоторые особенности изучаемых объектов, которые в обязательном порядке следует учитывать на подготовительной стадии издательско-полиграфического процесса. Для получения конкретного результата относительно качества учебных материалов полученное решающее правило следует реализовать на программном уровне.

В соответствии с требованиями к программным средствам был разработан программный продукт Readability analysis, предназначенный для автоматизированного контроля качества материалов подготовительной стадии (рис. 2).

На заключительном этапе работы проведена верификация с использованием 16 текстовых объектов (издательских оригиналов). Был проведен эксперимент на основе метода балльных оценок. После этого был произведен контроль качества выбранных объектов с помощью разработанной программы. Точность результатов — 94% (табл. 3).

Таким образом, на предприятиях издательско-полиграфической отрасли может использоваться схема контроля качества подготовительной стадии, представленная на рис. 3. Положительный эффект от использования данной модели в технологии допечатных процессов обусловлен, прежде всего, расширением экспертных методов контроля качества наиболее ранней стадии издательско-полиграфического производства.

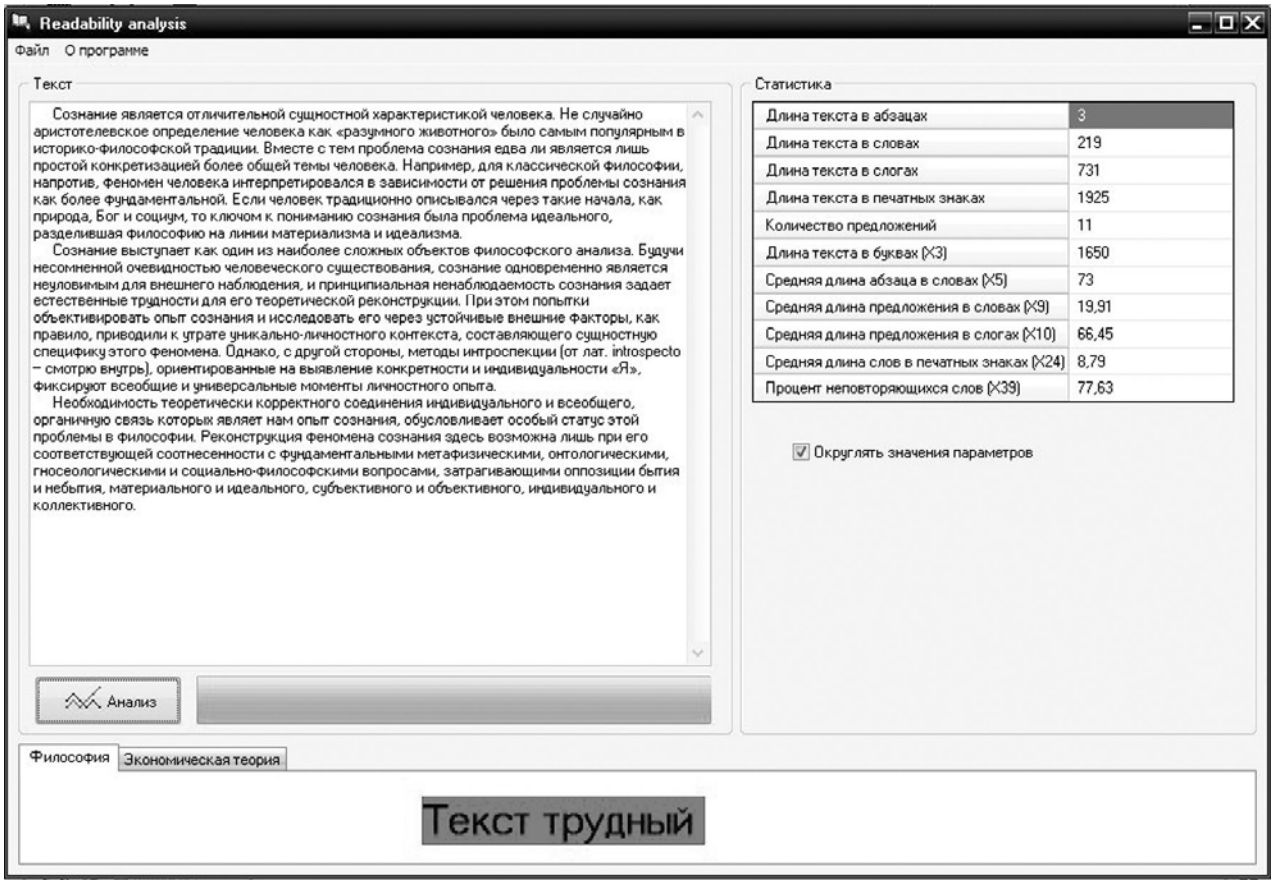

Рис. 2. Анализ текстового материала с помощью программного продукта

Таблица 3

Результаты верификации

№ объекта	Разбиение на классы в результате		Результат	№ объекта	Разбиение на классы в результате		Результат
	эксперимента	программной оценки			эксперимента	программной оценки	
1	1	1	Верно	9	0	0	Верно
2	0	0	Верно	10	0	0	Верно
3	1	1	Верно	*11	1	0	**Неверно**
4	1	1	Верно	12	1	1	Верно
5	0	0	Верно	13	1	1	Верно
6	0	0	Верно	14	1	1	Верно
7	1	1	Верно	15	0	0	Верно
8	1	1	Верно	16	1	1	Верно

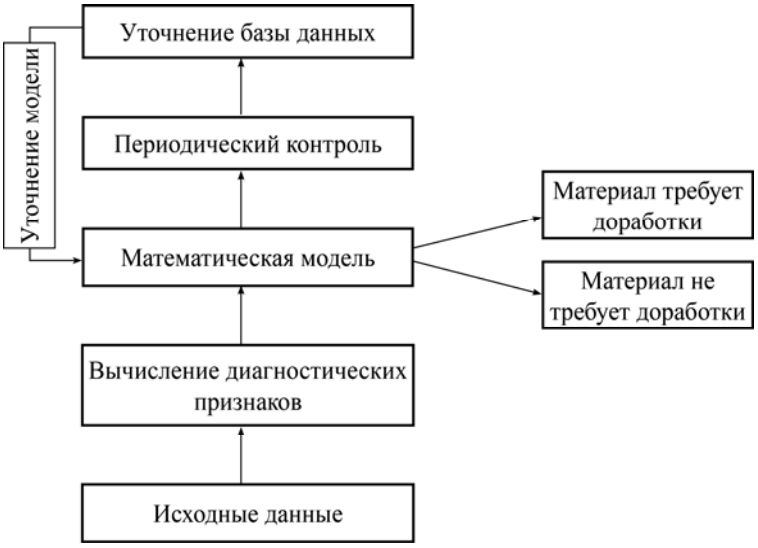

Рис. 3. Схема контроля качества учебных материалов на подготовительной стадии издательско-полиграфического производства

Выводы. *Основные научные результаты проведенного исследования:*

1. Предложен новый научно обоснованный подход для контроля качества материалов на допечатной стадии полиграфического производства на основе отбора наиболее информативных признаков и разработки модели контроля качества в виде решающего правила.

2. На основе методов многомерного статистического анализа (кластерный и факторный анализ, методы корреляционных плеяд и вроцлавской таксономии, многомерное шкалирование) выделены относительно однородные группы взаимосвязанных признаков и выявлена связь между данными группами и особенностями изучаемых объектов.

3. Впервые выявлены объективные диагностические показатели, которые оказывают наибольшее влияние на качество учебных материалов на подготовительной стадии издательско-полиграфического производства.

4. Впервые разработана модель контроля качества учебных материалов на подготовительной стадии издательско-полиграфического производства в виде решающего правила, что позволяет осуществить автоматизированную проверку качества на основе дискриминантного анализа. Точность классификации объектов составляет 83–92%. Разработанный в исследовании алгоритм в виде программного продукта интегрирован в существующие системы издательско-полиграфического комплекса для повышения эффективности процесса.

Рекомендации по практическому использованию результатов:

1. Программная реализация разработанного алгоритма контроля качества учебных материалов на допечатной стадии внедрена на следующих предприятиях: Управление редакционно-издательской работы БГУ, ООО «Современная школа», ООО «Харвест». Программа также внедрена в учебный процесс на кафедре редакционно-издательских технологий факультета издательского дела и полиграфии БГТУ. Внедрение результатов исследования, и в частности программа Readability analysis, несомненно, повысит качество учебных материалов, снизит

временные затраты путем отсеивания некачественных рукописей на стадии поступления их в книгоиздающие организации.

Программный продукт, в основу которого легла разработанная модель контроля качества, зарегистрирован в Национальном центре интеллектуальной собственности Республики Беларусь (свидетельство о регистрации компьютерной программы № 635; зарег. 14.02.2014).

2. Потенциально сфера применения программы Readability analysis не ограничивается учебными изданиями по философии и экономической теории и может быть расширена в случае научного подтверждения влияния выявленных факторов на трудность учебных текстов по другим тематическим разделам.

3. Перспективы дальнейшего развития данного научного направления заключаются в создании адекватных математических, экономических, технических и других формализованных моделей, разработке общей модели для контроля качества технологии допечатных процессов издательско-полиграфического производства и последующем ее совершенствовании.

Литература

1. Гринбаум О. Н., Мартыненко Г. Я., Фитиалов С. Я. Проект «ЛИНДА» — автоматизированная система обработки лингвостатистических данных // Прикладная лингвистика и автоматический анализ текста. Тарту: Изд-во ТГУ, 1988. С. 31–33.

2. Мартыненко Г. Я. Основы стилеметрии. Л.: Изд-во Ленингр. ун-та, 1988. 176 с.

3. Поддубный В. В., Шевелев О. Г., Фатыхов А. А. Сравнительный анализ эффективности алгоритмов распознавания авторства текстов по частотам переходов // Вестник Том. гос. ун-та. 2006. № 290. С. 232–234.

4. Пиотровский Р. Г. Текст, машина, человек. Л.: Наука, 1975. 327 с.

5. Хмелёв Д. В. Распознавание автора текста с использованием цепей А. А. Маркова // Вестник Моск. ун-та. Сер. 9, Филология. 2000. № 2. С. 115–126.

6. Chall J. S. Readability: an appraisal of research and application // Bureau of educational research monographs. Columbus, OH: Ohio State University Press, 1958. № 34. P. 58–68.

7. Шевелев О. Г. Разработка и исследование алгоритмов сравнения стилей текстовых произведений: автореф. дис. … канд. техн. наук по спец. 05.13.18 / Том. гос. ун-т. Томск, 2006. 20 с.

8. Хмелёв Д. В. Сложностной подход к задаче определения авторства текста // Русский язык: исторические судьбы и современность: тр. и материалы Междунар. конгр., Москва, 13–16 марта 2001 г. / Моск. гос. ун-т. М., 2001. С. 426–427.

9. Flesch R. Estimating the comprehension difficulty of magazine articles // Journal of general psychology. 1943. № 28. P. 63–80.

10. Мацковский М. С. Проблемы читабельности печатного материала // Смысловое восприятие речевого сообщения (в условиях массовой коммуникации). М., 1976. С. 126–142.

Поступила 20.03.2014