ПРИКЛАДНЫЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

УДК 004.93'14

М.А. Зильберглейт, М.М. Невдах, Ю.Ф. Шпаковский

ОЦЕНИВАНИЕ ТРУДНОСТИ ПОНИМАНИЯ УЧЕБНЫХ ТЕКСТОВ ДЛЯ ВЫСШЕЙ ШКОЛЫ

Проводятся эксперименты с использованием различных методик (методики дополнения, методов экспертных оценок и парных сравнений) для получения объективных критериев относительно трудности текстов. Выделяются и вычисляются значения 49 параметров учебных текстов. Осуществляется снижение признакового пространства методами многомерного статистического анализа. С помощью дискриминантного анализа разрабатывается решающее правило для оценивания трудности понимания учебных текстов для высшей школы.

Введение

Проблема качества учебных изданий является одной из центральных в отечественном книгоиздании и привлекает к себе внимание широкого круга исследователей. От повышения качества учебной литературы будет зависеть совершенствование профессиональной подготовки специалистов. В настоящее время уровень учебного материала в основном зависит от профессионализма автора и редактора. Очевидно, что данная оценка не всегда является объективной. В связи с этим создание надежных и общепринятых методов автоматизированной проверки трудности понимания учебного текста является крайне актуальной задачей.

Статистические методики анализа данных с поддержкой компьютерных технологий обладают огромным потенциалом в разрешении многих практических задач обработки текстовых массивов. Одной из областей анализа текста с точки зрения его доступности для читателя является читабельность, под которой следует понимать некоторую характеристику печатного материала, зависящую от всех элементов внутри данного материала, которые влияют на успешность его усвоения определенной группой читателей. Мерой такого успешного усвоения является то, насколько средний читатель интересующей нас группы понимает исследуемый материал, в какой мере скорость, с которой он его читает, приближается к оптимальной и, наконец, какой интерес представляет данный материал для этого среднего читателя.

В настоящее время отдельные аспекты читабельности с использованием современных информационных технологий привлекают специалистов из различных областей знаний, так как трудность учебных текстов волнует не только педагогов и родителей, но и самих учащихся. Исследования, посвященные автоматизированной оценке трудности учебных текстов (на примере белорусских изданий), не проводились.

Целью настоящей работы является автоматизированная оценка трудности учебных текстов по философии и экономической теории для высшей школы. В решении этой задачи можно выделить несколько этапов: 1) нахождение и реализация методов для определения трудности понимания различных текстов данной группой лиц; 2) выбор формальных характеристик текста (и только тех, которые поддаются точному измерению); 3) создание автоматизированной системы, которая бы на основе ответов испытуемых, полученных экспериментальным путем, предсказывала понятность текста для будущих читателей.

1. Вычисление критериев трудности понимания учебных текстов

Экспериментальным материалом послужили учебные издания для вузов по философии и экономической теории [1–12]. Всего отобрано 48 отрывков длиной 1800–2000 печатных знаков. Выбор данной величины обусловлен тем, что, начиная с объема в 1800 печатных знаков, статистические характеристики текста становятся относительно постоянными [13].

На ***первом этапе*** проведены эксперименты с использованием различных методик. С этой целью были проанализированы основные методы определения трудности понимания текста [14]: постановка вопросов к тексту, сводка основного содержания текста, составление плана или схем текста, интонирование, пересказ и др. В нашем исследовании применялись наиболее надежные методы: методика дополнения и метод экспертных оценок. Впервые для оценки трудности понимания учебного материала использовался метод парных сравнений.

Методика дополнения – это заполнение пропусков в тексте, в котором слова через определенный интервал заменены точками. Плюсы данной методики состоят в том, что пропускается всегда только одно слово и не по усмотрению исследователя, а по строгому правилу. В текстах на основе результатов предварительного эксперимента пропускалось каждое седьмое слово.

Суть *экспертных оценок* трудности текста заключалась в следующем: после прочтения отрывка испытуемым предлагалось оценить его трудность по семибалльной шкале: 1 – сверхлегкий текст, 2 – очень легкий, 3 – легкий, 4 – текст со средней трудностью, 5 – трудный, 6 – очень трудный, 7 – сверхтрудный текст. Для того чтобы исключить поверхностное знакомство испытуемых с текстом и возможное искажение результатов при оценке его трудности, студентам перед оценкой трудности понимания текста по шкале предлагалось выписать несколько ключевых слов и выразить основное содержание отрывка одним предложением. При проведении методики дополнения и экспертных оценок фиксировалось также время работы с текстом.

Суть *метода парных сравнений* заключалась в том, что каждому испытуемому предлагался набор текстов, размещенных парами, и после прочтения студент должен был указать, какой из отрывков обладает заданным признаком (в нашем случае – какой отрывок воспринимается легче). Оценка каждого текста производилась путем сравнения с каждым другим текстом того же набора. Так как у нас в наборе имелось 24 отрывка по философии и столько же по экономической теории, следовательно, по одному предмету было составлено 276 пар. За один этап эксперимента студенту предъявлялось восемь пар текстов. Такое количество не вызывало утомления у испытуемого.

Оценка трудности учебных текстов проводилась среди студентов (75 человек) старших курсов Белорусского государственного технологического университета.

Обработка и анализ результатов экспериментов позволили выявить информацию относительно трудности понимания учебного материала. На основании полученных данных найдены пять объективных критериев, определяющих трудность текста: процент правильно заполненных пропусков (Y_1); относительное время работы с текстом (Y_2) – с использованием методики дополнения; средняя оценка трудности восприятия текста (Y_3); относительное время работы с текстом (Y_4) – с использованием экспертных оценок; ранг текста (Y_5). Результаты экспериментов были сведены в таблицу, фрагмент которой представлен ниже.

Таблица 1

Критерии трудности учебных текстов для высшей школы

Номер теста	Методика дополнения правильно заполненных пропусков, %	Относительное время работы с текстом (по методике дополнения)	Экспертные оценки испытуемых	Относительное время работы с текстом (по экспертным оценкам испытуемых)	Метод парных сравнений (ранг)
1	78,82	38,55	4,21	0,017	6
2	77,11	30,74	3,51	0,014	22
3	71,02	39,29	3,31	0,020	12
4	53,69	35,44	3,63	0,024	14
…	…	…	…	…	…
48	55,46	28,40	4,57	0,022	3

Для каждого показателя найдена середина диапазона всех полученных значений, в соответствии с которой производилось разбиение текстов на две группы: трудные (0), легкие (1). В итоге было получено разбиение текстов на группы по выделенным пяти показателям (табл. 2).

Таблица 2

Разбиение текстов на группы в соответствии с субъективной оценкой трудности текста

Номер теста	Методика дополнения правильно заполненных пропусков, %	Относительное время работы с текстом (по методике дополнения)	Экспертные оценки испытуемых	Относительное время работы с текстом (по экспертным оценкам испытуемых)	Метод парных сравнений (ранг)
1	1	0	0	1	0
2	1	1	1	1	1
3	1	0	1	0	0
4	0	1	1	0	1
…	…	…	…	…	…
48	0	1	0	0	0

2. Изучение текстовых параметров методами многомерного статистического анализа

Объективная трудность учебных текстов определялась путем анализа компонентов сложности текстов. Для этого на ***втором этапе*** были выделены и вычислены значения 49 параметров учебных текстов по философии и экономической теории: 1) длина текста в абзацах; 2) длина текста в словах; 3) длина текста в буквах; 4) средняя длина абзаца во фразах; 5) средняя длина абзаца в словах; 6) средняя длина абзаца в буквах; 7) средняя длина абзаца в печатных знаках; 8) средняя длина предложения во фразах; 9) средняя длина предложения в словах; 10) средняя длина предложения в слогах; 11) средняя длина предложения в буквах; 12) средняя длина предложения в печатных знаках; 13) средняя длина самостоятельного предложения во фразах; 14) средняя длина самостоятельного предложения в словах; 15) средняя длина самостоятельного предложения в слогах; 16) средняя длина самостоятельного предложения в буквах; 17) средняя длина самостоятельного предложения в печатных знаках; 18) средняя длина фразы в словах; 19) средняя длина фразы в слогах; 20) средняя длина фразы в буквах; 21) средняя длина фразы в печатных знаках; 22) средняя длина слов в слогах; 23) средняя длина слов в буквах; 24) средняя длина слов в печатных знаках; 25) средняя длина слов по Деверу; 26) процент слов длиной в 5 букв и больше; 27) процент слов длиной в 6 букв и больше; 28) процент слов длиной в 7 букв и больше; 29) процент слов длиной в 8 букв и больше; 30) процент слов длиной в 9 букв и больше; 31) процент слов длиной в 10 букв и больше; 32) процент слов длиной в 11 букв и больше; 33) процент слов длиной в 12 букв и больше; 34) процент слов длиной в 13 букв и больше; 35) процент слов в 3 слога и больше; 36) процент слов в 4 слога и больше; 37) процент слов в 5 слогов и больше; 38) процент слов в 6 слогов и больше; 39) процент неповторяющихся слов; 40) средняя частота повторения слова; 41) процент неповторяющихся существительных; 42) процент повторяющихся существительных; 43) процент конкретных существительных; 44) процент абстрактных существительных; 45) процент прилагательных; 46) процент глаголов; 47) процент сложных предложений; 48) процент простых предложений; 49) процент придаточных предложений среди фраз.

Под термином «фраза» в данной статье понимается отрезок текста, в котором содержится одна предикативная связь. Исходя из этого к фразе относятся простое предложение, части сложносочиненного предложения, главное и придаточное предложения в сложноподчиненном. Самостоятельным предложением считаются простые предложения, части сложносочиненного предложения и сложноподчиненное в целом. Средняя длина слов по Деверу рассчитывалась делением общего количества знаков с пробелами на число знаков без пробелов.

Очевидно, что использование большого количества параметров текста является неэффективным по ряду причин: сильная взаимосвязанность признаков, которая приводит к дублированию информации; неинформативность признаков, мало меняющихся при переходе от одного объекта к другому; возможность агрегирования по некоторым признакам. Однако ничем не оправданное уменьшение числа переменных может привести к потере точности экспериментов.

Для снижения признакового пространства были использованы кластерный и факторный анализы, метод корреляционных плеяд и вроцлавской таксономии, многомерное шкалирование.

Так как характеристики текста измерялись в различных единицах, данные были стандартизированы. Для этого применялась нормализация, приводящая все переменные к стандартной *z*-шкале. Для анализа данных и проведения статистического анализа использован пакет SPSS.

Кластерный анализ представляет собой многомерную статистическую процедуру, которая выполняет сбор данных, содержащих информацию о выборке объектов, и затем упорядочивает объекты в сравнительно однородные группы.

В данном исследовании для анализа данных в качестве критерия для определения подобия групп использовались следующие меры сходства: расстояние Евклида; квадрат расстояния Евклида; косинус угла; коэффициент корреляции; неравенство Чебышева; расстояние Минковского; манхэттенское расстояние.

Для кластеризации выделенных характеристик текста использовались следующие основные алгоритмы метода кластерного анализа: метод простого среднего (межгрупповое связывание), метод группового среднего (внутригрупповое связывание), метод ближнего соседа (одиночное связывание), метод дальнего соседа (полное связывание), невзвешенный центроидный метод (центроидная кластеризация), взвешенный центроидный метод (центральное связывание), метод Варда. Количество кластеров по каждому алгоритму варьировалось от 3 до 10. После выбора всех соответствующих параметров получена информация по формированию кластеров: порядок объединения кластеров, расстояние между ними, а также принадлежность характеристик текста к тому или иному кластеру. Пример анализа формирования кластеров для учебных текстов представлен в табл. 3.

Таблица 3

Кластеризация на примере использования алгоритма «Метод Варда», основанного на расстоянии Евклида

Признак	10 кластеров	9 кластеров	8 кластеров	7 кластеров	6 кластеров	5 кластеров	4 кластера	3 кластера
1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	1
3	3	3	3	3	3	3	3	2
4	1	1	1	1	1	1	1	1
…	…	…	…	…	…	…	…	…
49	8	7	7	6	4	4	2	1

Выводимые результаты для наглядности были представлены и в виде дендрограмм, которые позволяют не только перейти к любому признаку на любом уровне кластеризации, но и дают возможность судить о том, каково расстояние между кластерами или признаками на каждом из уровней (рис. 1).

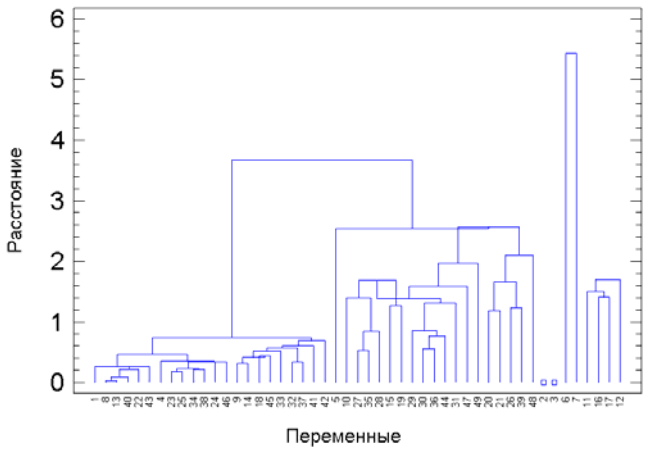

Рис. 1. Дендрограмма по центроидному методу на основе манхэттенского расстояния для пяти кластеров

В результате анализа данных о влиянии исследуемых характеристик текста с использованием всех известных алгоритмов и мер сходства были получены 784 дендрограммы, которые отражают кластеризацию переменных в условные группы.

В применении процедур кластерного анализа немаловажным аспектом является устойчивость структуры кластеров, отражающая реальную объективность классификации. В качестве методов проверки устойчивости могут быть использованы бутстреп-метод, предложенный Б. Эфроном в 1977 г., методы «складного ножа» и «скользящего контроля» [15]. Одним из наиболее простых и эффективных способов проверки устойчивости результатов является метод сравнения результатов, полученных для различных алгоритмов кластеризации, который и использовался в данной работе. Для этого все данные для наглядности были объединены в сводные таблицы, в которых четко прослеживаются особенности применения различных алгоритмов кластерного анализа, использующих разные меры сходства. Результаты формирования кластеров для текстов по философии согласуются практически по всем алгоритмам. Незначительно отличаются данные по методу Варда и центроидной кластеризации. Сравнение результатов с применением различных мер сходства показало, что наблюдаются заметные различия лишь в данных, полученных методами измерения близости и основанных на косинусах векторов значений и корреляции векторов значений.

Для текстов по экономической теории результаты формирования кластеров по различным алгоритмам отличаются только по методу Варда. При использовании различных мер сходства наблюдаются заметные различия лишь в данных, полученных методами измерения близости и основанных на корреляции векторов значений и манхэттенском расстоянии.

Проведенный кластерный анализ для учебных текстов показал, что целесообразно выделить следующие группы признаков: философия – 1, 4, 8, 13, 18, 22–25, 33, 34, 38, 40, 43, 45, 46; 2; 3; 5; 6, 7; 9, 14, 19, 30–32, 36, 37, 41, 42, 44; 10, 15, 27–29, 35, 47, 49; 11, 12, 16, 17; 20, 21, 26, 39, 48 (9 групп); экономика – 1, 4, 8, 13, 22–25, 40, 42–44, 46; 2, 9, 14, 18; 3, 39, 45, 48; 5–7; 10–12, 16, 17; 15, 19–21; 26–30, 35–37; 31–34, 38, 41; 47, 49 (9 групп). Для последующей обработки достаточно пользоваться одним признаком из каждой группы.

Снижение размерности набора переменных в методах *факторного анализа* базируется в основном на взаимной коррелированности исходных признаков. В связи с этим первый этап исследования заключался в вычислении корреляционной матрицы.

При изучении экспериментальных данных было установлено, что первые три фактора объясняют около 74 % разброса дисперсии для текстов по философии, около 64 % – для текстов по экономической теории.

Так как факторный анализ является методом сокращения числа переменных, возникает вопрос, какие из факторов следует оставить для дальнейшей обработки. Исследователи рекомендуют руководствоваться здравым смыслом и оставлять только те факторы, которые имеют понятную или логическую интерпретацию. Однако установить заранее назначение каждого фактора не всегда представляется возможным, поэтому для начала были использованы формальные критерии: критерий Кайзера и критерий «каменистой осыпи» Р. Кэтелла.

Первый критерий, как правило, сохраняет слишком много факторов, в то время как второй – слишком мало, поэтому решение об оптимальном количестве факторов можно принять только после их вращения и интерпретации.

Целью вращения факторов является получение простой структуры, которой соответствует большое значение нагрузки каждой переменной только по одному фактору и малое – по всем остальным факторам. Нагрузка (значение лежит в пределах от –1 до 1) отражает связь между переменной и фактором. В работе использовались ортогональные методы вращения: варимакс, квартимакс и эквимакс. В результате были получены матрицы нагрузок для переменных.

Изучение результатов с использованием всех методов факторного анализа и методов вращения позволило выявить, как признаки распределились между факторами (табл. 4 и 5).

Как видно из табл. 4 и 5, факторы по всем методам вращения для текстов по философии и экономической теории практически идентичны. Для более ясного представления о распределении переменных использовались диаграммы рассеяния (рис. 2).

Таблица 4

Распределение характеристик текстов по философии с использованием методов факторного анализа и вращения

Методы вращения	Методы факторного анализа								
	метод главных факторов			центроидный метод			метод главных компонент		
	фактор 1	фактор 2	фактор 3	фактор 1	фактор 2	фактор 3	фактор 1	фактор 2	фактор 3
Варимакс	22, 23, 25–40	1, 5–14, 16, 17, 47, 48	18, 41, 42	22, 23, 25–40	1, 5, 14, 16, 17, 47, 48	18, 41, 42	22, 23, 25–40	1, 5, 14, 16, 17, 47, 48	18, 41, 42
Квартимакс	19–23, 25–40	1, 5–7, 9–17, 47, 48	41, 42	19–23, 25–40	1, 5–7, 9–17, 47, 48	41	19–23, 25–40	1, 5–7, 9–17, 47, 48	41, 42
Эквимакс	19–23, 25–40	1, 5–7, 9–17, 47, 48	41, 42	19–23, 25–40	1, 5–7, 9–17, 47, 48	41	19–23, 25–40	1, 5–7, 9–17, 47, 48	41, 42

Таблица 5

Распределение характеристик текстов по экономической теории с использованием методов факторного анализа и вращения

Методы вращения	Методы факторного анализа											
	метод главных факторов				центроидный метод				метод главных компонент			
	фактор 1	фактор 2	фактор 3	фактор 4	фактор 1	фактор 2	фактор 3	фактор 4	фактор 1	фактор 2	фактор 3	фактор 4
Варимакс	2, 22–25, 27–37	9–12, 14, 16–18, 20, 21	13, 15, 19, 42–44	1, 4–7	2, 22–37	9–12, 14, 16–18, 20, 21	13, 15, 19, 42–44	1, 4–7	2, 22–38	9–12, 14, 16–18, 20, 21	13, 15, 19, 42–44	1, 4–7
Квартимакс	2, 22–25, 27–38	9–12, 14, 16–18, 20, 21	13, 15, 19, 42–44	1, 4–7	2, 22–37	9–12, 14, 16–18, 20, 21	13, 15, 19, 42–44	1, 4–7	2, 22–38	9–12, 14, 16–18, 20, 21	13, 15, 19, 42–44	1, 4–7
Эквимакс	2, 22–25, 27–38	9–12, 14, 16–18, 20, 21	13, 15, 19, 42–44	1, 4–7	2, 22–37	9–12, 14, 16–18, 20, 21	13, 15, 19, 42–44	1, 4–7	2, 22–38	9–12, 14, 16–18, 20, 21	13, 15, 19, 42–44	1, 4–7

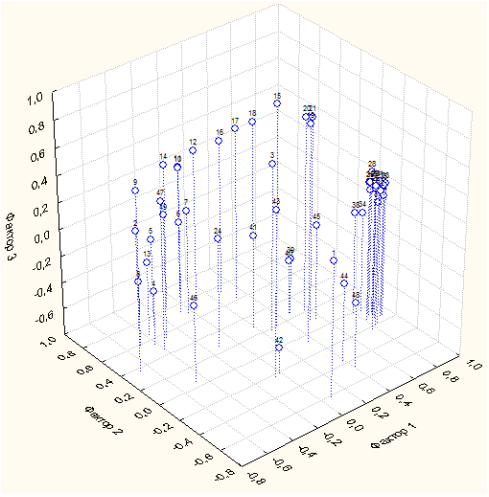

Рис. 2. Диаграмма рассеяния признаков для метода главных факторов

Результаты, полученные методом главных факторов, центроидным методом и методом главных компонент, позволяют выделить следующие группы признаков: философия – 1, 44, 48; 2, 3, 5–17, 24, 47, 49; 4, 46; 18–21, 41; 22, 23, 25–40; 42, 43 (6 групп); экономика – 1, 4, 8, 13, 22–25, 40, 42–44, 46; 2, 9, 14, 18; 3, 39, 45, 48; 5–7; 10–12, 16, 17; 15, 19–21; 26–38, 41; 47, 49 (8 групп).

Для снижения признакового пространства использовался и метод *корреляционных плеяд*. Выделение корреляционных плеяд осуществлялось следующим образом: признаки упорядочивались и рассматривались только те коэффициенты корреляции, которые соответствуют связям между элементами в упорядоченной системе.

Упорядочение производилось на основании принципа максимального корреляционного пути. Для удобства построения графа были составлены упорядоченные корреляционные матрицы (табл. 6).

Таблица 6

Фрагмент упорядоченной корреляционной матрицы исходных признаков текстов по философии

Номер признака	47	48	8	13	4	5	6	…	43
47	1	1,000	0,878	0,818	0,624	0,468	0,307	…	0,225
48		1	0,878	0,818	0,624	0,467	0,306	…	0,226
8			1	0,764	0,733	0,525	0,336	…	0,228
13				1	0,530	0,474	0,360	…	0,174
4					1	0,796	0,650	…	0,346
5						1	0,953	…	0,308
6							1	…	0,273
…								…	…
43									1

На основании упорядочения признаков были построены графы, которые представляют собой кратчайший незамкнутый путь (рис. 3).

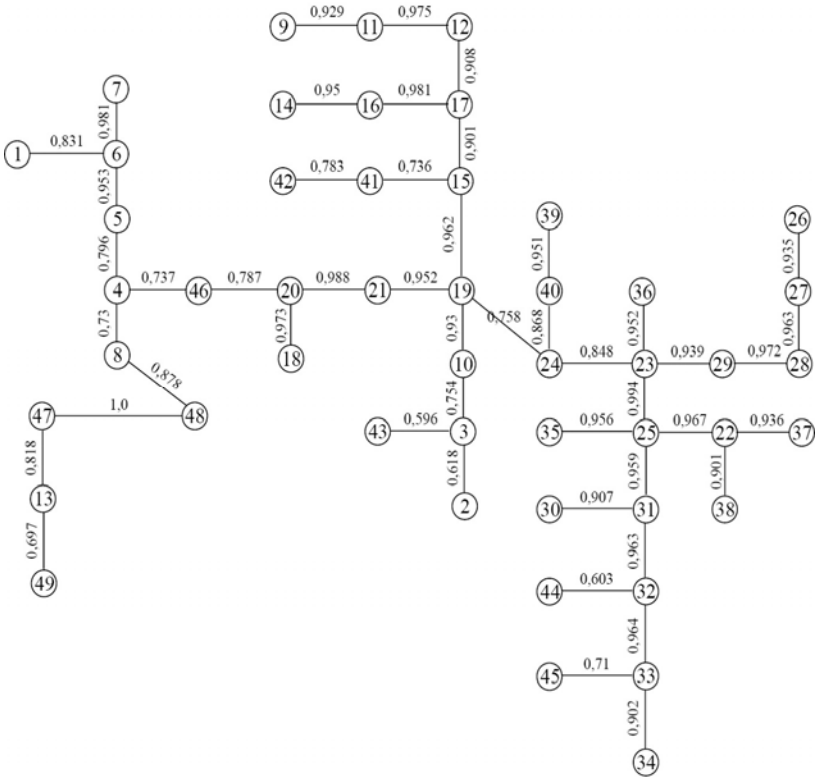

Рис. 3. Граф максимального корреляционного пути (для текстов по философии)

В графах соединены все исследуемые параметры текстов. Если задать определенное пороговое значение коэффициента корреляции (r_0), то полученные графы можно разбить на подграфы (плеяды), проводя разрыв между признаками со значением сопряженности, меньшим r_0.

Используя прямое (z) и обратное (z^{-1}) преобразования Фишера, было определено r_0 для заданного объема выборки. Исходя из поставленной цели и анализа корреляционной матрицы исследуемых характеристик текстов был задан пороговый коэффициент корреляции $r \geq 0{,}9$, что позволило выявить наиболее связанные друг с другом признаки. Исходный граф распался на пять подграфов для текстов по философии и шесть подграфов – для текстов по экономической теории. Признаки, не вошедшие в выделенные группы, требуют дальнейшего исследования. Использование метода корреляционных плеяд позволило выделить следующие группы близких параметров текста: философия – 1; 2; 3; 4; 5-7; 8; 9–12, 14–21; 22, 23, 25–38; 24; 39, 40; 41; 42; 43; 44; 45; 46; 47, 48; 49 (18 групп); экономика – 1; 2; 3; 4; 5–7; 8; 9, 11, 12, 14, 16–18, 20, 21; 10; 13; 15, 19; 22–25, 27–33, 35–37; 26; 34; 38; 39, 40; 41; 42; 43; 44; 45; 46; 47, 48; 49 (23 группы).

С помощью метода *вроцлавской таксономии* было получено нелинейное упорядочение изучаемых элементов текста. С целью построения дендрита были вычислены матрицы расстояний (на основе расстояния Евклида) между изучаемыми характеристиками учебных текстов (табл. 7).

Таблица 7

Фрагмент матрицы расстояний исходных признаков

Номер признака	1	2	3	4	5	6	7	8	9	…	49
1	0	4,075	25,833	0,090	1,384	8,828	9,165	0,037	0,275	…	0,699
2	4,075	0	21,782	4,002	2,770	4,978	5,307	4,103	3,818	…	3,444
3	25,833	21,782	0	25,761	24,521	17,413	17,105	25,861	25,578	…	25,193
4	0,090	4,002	25,761	0	1,301	8,747	9,083	0,107	0,200	…	0,631
5	1,384	2,770	24,521	1,301	0	7,454	7,789	1,406	1,119	…	0,797
6	8,828	4,978	17,413	8,747	7,454	0	0,345	8,851	8,563	…	8,188
7	9,165	5,307	17,105	9,083	7,789	0,345	0	9,187	8,899	…	8,524
…	…	…	…	…	…	…	…	…	…	…	…
49	0,699	3,444	25,193	0,631	0,797	8,188	8,524	0,722	0,454	…	0

Далее из составленных матриц расстояний между признаками были выбраны единицы с близкими значениями. В результате для текстов по философии были получены следующие пары признаков с близкими значениями: 1–22, 2–12, 3–7, 4–25, 5–39, 6–7, 7–6, 8–13, 9–14, 10–15, 11–12, 12–11, 13–8, 14–9, 15–10, 16–17, 17–16, 18–46, 19–30, 20–21, 21–20, 22–1, 23–25, 24–46, 25–23, 26–39, 27–35, 28–35, 29–36, 30–19, 31–37, 32–37, 33–34, 34–38, 35–27, 36–44, 37–32, 38–34, 39–26, 40–13, 41–31, 42–45, 43–22, 44–36, 45–18, 46–18, 47–10, 48–26, 49–15. Необходимо отметить, что некоторые пары повторяются дважды, например 6–7 и 7–6. Так как при построении дендрита очередность установления связей не имеет значения, одно из повторяющихся сочетаний следует исключить. Далее были найдены пары с общим признаком, которые затем объединялись друг с другом. Например, пары 43–22 и 22–1 образовали цепочку 43–22–1. В результате было получено 16 отдельных конструкций, называемых скоплениями первого порядка: 43–22–1, 2–12–11, 3–7–6, 4–25–23, 5–39–26–48, 8–13–40, 9–14, 49–15–10–47, 16–17, 42–45–18–46–24, 19–30, 20–21, 27–35–28, 29–36–44, 41–31–37–32, 33–34–38.

Полученные скопления не удовлетворяют основному условию дендрита, а именно они не связаны в единое целое. Для достижения этой цели было выбрано наименьшее расстояние между единицами, входящими в различные скопления первого порядка. В результате были получены скопления второго порядка. Объединение признаков в скопления третьего, четвертого, n-го порядков происходило до тех пор, пока любые две точки исследуемого множества параметров не оказались связанными друг с другом.

Исходя из поставленной цели и анализа дендрита, была определена максимальная величина расстояния между признаками, равная 0,08. Исходный дендрит распался на следующих

семь наиболее связанных друг с другом групп признаков: 1) 1, 4, 8, 13, 18, 22, 23, 25, 33, 34, 38, 40, 43, 45, 46; 2) 9 и 14; 3) 11 и 12; 4) 16 и 17; 5) 20 и 21; 6) 27 и 35; 7) 31, 32 и 37.

Для текстов по экономической теории все шаги были повторены. Максимальная величина расстояния между признаками была определена равной 0,15. Исходный дендрит распался на пять наиболее связанных друг с другом групп признаков.

Использование метода вроцлавской таксономии позволило выделить следующие группы признаков: философия – 1, 4, 8, 13, 18, 22, 23, 25, 33, 34, 38, 40, 43, 45, 46; 2; 3; 5; 6; 7; 9, 14; 10; 11, 12; 15; 16, 17; 19; 20, 21; 24; 26; 27, 35; 28; 29; 30; 31, 32, 37; 36; 39; 41; 42; 44; 47; 48; 49 (28 групп); экономика – 1, 4, 8, 9, 13, 14, 18, 22–25, 34, 38, 40–43, 45, 46; 2; 3; 5; 6; 7; 10; 11; 12; 15, 19; 16; 17; 20; 21; 26, 39; 27, 28, 35; 29–33, 36, 37; 44; 47; 48; 49 (21 группа).

В данной работе использовалось и *многомерное шкалирование*, основная задача которого заключалась в преобразовании исходной матрицы 49×49 в гораздо более простую матрицу 49×2 и визуальным представлением ее в виде диаграммы (рис. 4).

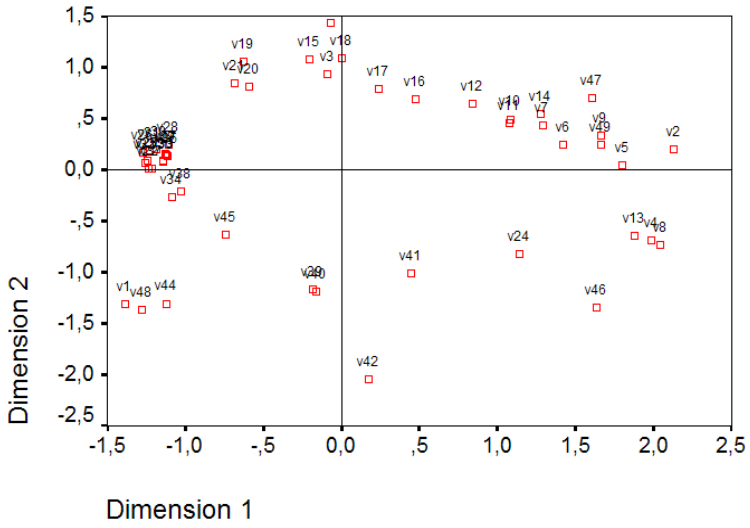

Рис. 4. Расположение точек на основе квадрата расстояния Евклида (для текстов по философии)

Смысл каждой из шкал не имеет значения, главным является взаимное расположение точек. После расположения точек в заданном пространстве для всей модели в многомерном шкалировании были вычислены стресс и коэффициент R^2. Наилучшей моделью для текстов по философии (stress = 0,210, R^2 = 0,856) стала модель, полученная с использованием меры сходства, которая основана на неравенстве Чебышева; по экономической теории (stress = 0,230, R^2 = 0,763) – модель, полученная с использованием квадрата расстояния Евклида. На их основе были получены следующие группы признаков: философия – 1; 2, 4–14, 47, 49; 3, 15–18; 19–21, 45; 22, 23, 25–38; 24; 39, 40; 41; 42, 43; 44, 48; 46 (11 групп); экономика – 1, 39, 40, 46; 2; 3, 4, 9–12, 14, 16, 17, 47, 49; 5–7, 18; 8, 13,42, 44; 15, 19–21, 43; 22–38, 41, 45, 48 (7 групп).

Для дальнейшего изучения характеристик текста важнейшей задачей является выделение наиболее информативного признака из каждой полученной группы. В данной работе для оценки информативности признаков в качестве информационной использовалась мера $J(1, 2)$ расхождения между статистическими распределениями 1 и 2. Для дискретных распределений эта мера вычисляется по формуле

$$J(x_i/A_1, x_i/A_2) = \sum_j J(x_i/A_1, x_i/A_2) = \sum_j \lg \frac{P(x_{ij}/A_1)}{P(x_{ij}/A_2)} \left[P(x_{ij}/A_1) - P(x_{ij}/A_2)\right],$$

где j – номер диапазона признака x_i; i – номер признака; A_1 и A_2 – классы, которым может принадлежать рассматриваемый объект; $P(x_{ij}/A_1)$ и $P(x_{ij}/A_2)$ – вероятности попадания объекта, принадлежащего к A_1 или к A_2, в диапазон j признака x_i.

По формуле, приведенной выше, были вычислены информационные меры каждого из 49 признаков, а затем отобраны те из них, которые обладают наибольшей информативностью среди признаков своей группы. В результате число признаков было сокращено до возможного минимума.

3. Разработка решающего правила для оценивания трудности понимания учебных текстов для высшей школы

Для дальнейшего исследования характеристик текста и их влияния на понятность учебного материала использовался дискриминантный анализ. Критериями для включения и исключения факторов являются пороговые значения F-критерия. При проведении дискриминантного анализа использовались значения $F = 3{,}84$ для включения факторов (что соответствует величине уровня значимости p, равной 0,05) и $F = 2{,}71$ для исключения факторов (что соответствует величине уровня значимости p, равной 0,1).

В работах И. Лорджа [16] и Р. Флеша [17] доказывается тот факт, что корреляция между факторами, влияющими на трудность понимания текста, настолько велика, что только некоторые из них необходимы для использования в качестве достоверных факторов трудности текста. Наибольшей популярностью в США пользуется формула читабельности Р. Флеша, в которую входят всего два параметра: средняя длина предложения в словах и средняя длина слова в слогах.

Таким образом, при анализе дискриминантных функций в учет принимались только функции, у которых процент точности классификации максимальный, а количество переменных при этом минимальное.

В текстах по философии в дискриминантных функциях чаще других фигурировали следующие признаки (в порядке убывания): 3 (длина текста в буквах), 24 (средняя длина слов в печатных знаках), 9 (средняя длина предложения в словах), 40 (средняя частота повторения слова), 1 (длина текста в абзацах), 43 (процент конкретных существительных), 48 (процент простых предложений); в текстах по экономической теории: 39 (процент неповторяющихся слов), 47 (процент сложных предложений), 48 (процент простых предложений), 15 (средняя длина самостоятельного предложения в слогах), 42 (процент повторяющихся существительных), 41 (процент неповторяющихся существительных).

Выделенные признаки позволяют сделать важный вывод относительно факторов трудности текста. Они связаны прежде всего с объемом текста (признак 3), с длиной слов и предложений (признаки 9, 15, 24), со сложностью организации текста (признаки 1, 47, 48), богатством словаря и абстрактностью изложения материала (признаки 39, 40–43). Эти выводы согласуются с исследованиями Я.А. Микка [18].

Анализ результатов дискриминантного анализа показал, что для учебных текстов по философии наилучшими являются следующие дискриминантные функции:

$$F_1 = -53{,}06 + 15{,}10X_{24} + 0{,}83X_9 - 0{,}01X_3;$$

$$F_2 = -42{,}72 + 8{,}66X_{24} + 0{,}55X_9 + 0{,}01X_3.$$

Точность классификации при данном наборе дискриминантных переменных составляет 91,6 % (22 из 24 правильных предсказаний в отношении известных объектов).

Для учебных текстов по экономической теории наилучшими являются следующие дискриминантные функции:

$$F_1 = -92{,}96 + 2{,}62X_{39} - 0{,}03X_{47};$$

$$F_2 = -111{,}93 + 2{,}96X_{39} - 0{,}20X_{47}.$$

Точность классификации при данном наборе дискриминантных переменных составляет 87,5 % (21 из 24 правильных предсказаний в отношении известных объектов).

Для автоматизированной оценки рукописи на основе полученных дискриминантных функций создана программа Readability analysis, предназначенная для автоматизации оценки

трудности учебных текстов для студентов вузов. Программа написана на языке Delphi и включает в себя три подпрограммы: «Расчет текстовых параметров», «Вычисление дискриминантных функций», «Вывод результатов». Новизна разработанного алгоритма программы (рис. 5) заключается в том, что при оценке трудности текста он учитывает только статистически значимые (наиболее информативные) параметры. Кроме того, в основе алгоритма лежит не уравнение регрессии, которое используется в большинстве известных алгоритмов, а решающее правило в виде дискриминантных функций, которое позволяет объективно оценить трудность учебного текста.

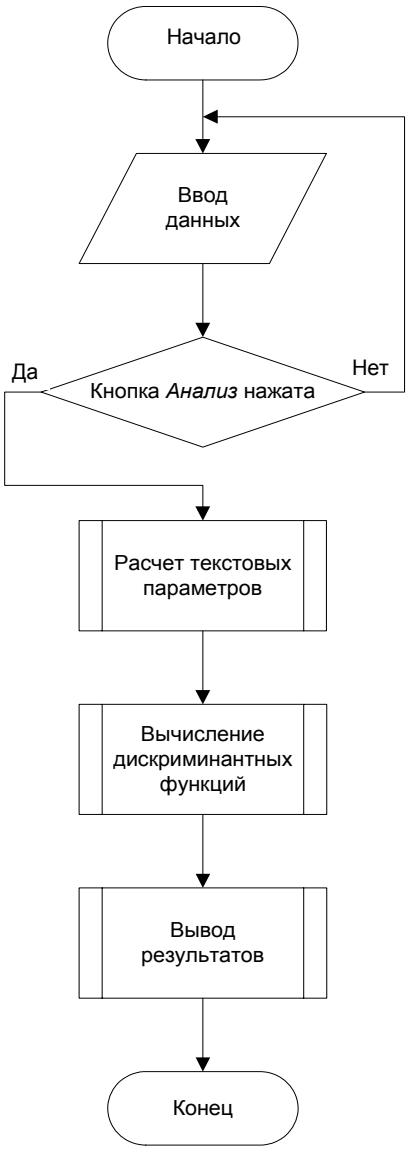

Рис. 5. Укрупненная блок-схема программы для оценки трудности понимания учебных текстов Readability analysis

Программа быстро справляется с обработкой больших массивов текстовой информации. Время оценки текста – от нескольких секунд до двух минут.

Практическая значимость программы Readability analysis связана с тем, что она может быть использована в редакционно-издательской деятельности при подготовке учебной литературы для высшей школы. Анализ трудности текста на стадии его подготовки и дальнейшее усовершенствование материала позволят привести уровень сложности учебного текста в соответствие со способностями читателей.

Заключение

Научная новизна предложенной методики оценки трудности учебного текста заключается в следующем:

– при разработке методики учитываются уровень подготовленности и способности обучающихся (студентов) путем проведения психолингвистических экспериментов, впервые для субъективной оценки трудности учебного текста использован метод парных сравнений;

– на основе анализа компонентов сложности текста выделены основные статистические характеристики учебных текстов для высшей школы;

– при реализации методики используется алгоритм, который учитывает только статистически значимые (наиболее информативные) параметры. Кроме того, в основе алгоритма лежит решающее правило, которое позволяет объективно оценить трудность учебного текста.

Результаты исследования дают возможность продолжить автоматизацию редакционно-издательского процесса. Полная или частичная замена человека специализированной системой позволит добиться не только невозможного для человека быстродействия, но и необходимого качества изданий благодаря объективной оценке трудности текста на основе его информационных характеристик, полученных с учетом восприятия читателей.

Результаты проведенного исследования прошли апробацию и внедрены в деятельность ряда государственных и частных издательств («Современная школа», «Вышэйшая школа», «Харвест», Управление редакционно-издательской работы БГУ).

Список литературы

1. Волчек, Е.З. Философия : учеб. пособие с хрестоматийными извлечениями / Е.З. Волчек. – Минск : Интерпресссервис, Экоперспектива, 2003. – 544 с.

2. Спиркин, А.Г. Философия : учебник для студентов высших учебных заведений / А.Г. Спиркин. – 2-е изд. – М. : Гардарики, 2004. – 736 с.

3. Философия : учебное пособие для студентов высших учебных заведений / В.С. Степин [и др.] ; под общ. ред. Я.С. Яскевич. – Минск : РИВШ, 2006. – 624 с.

4. Философия : учебное пособие для студентов высших учебных заведений / Ю.А. Харин [и др.] ; под общ. ред. Ю.А. Харина. – Минск : ТетраСистемс, 2006. – 448 с.

5. Сажина, М.А. Основы экономической теории : учебное пособие для неэкономических специальностей вузов / М.А. Сажина, Г.Г. Чибриков ; отв. ред. и руководитель авт. коллектива П.В. Савченко. – М. : Экономика, 1995. – 368 с.

6. Экономическая теория : учебник / Н.И. Базылев [и др.] ; под общ. ред. Н.И. Базылева, С.П. Гурко. – Минск : Экоперспектива, 1997. – 368 с.

7. Экономическая теория: учебник для студентов вузов ; под ред. В.Д. Камаева. – 6-е изд., перераб. и доп. – М. : ВЛАДОС, 2001. – 640 с.

8. Экономическая теория : учебное пособие / Л.Н. Давыденко [и др.] ; под общ. ред. Л.Н. Давыденко. – Минск : Вышэйшая школа, 2002. – 366 с.

9. Кажуро, Н.Я. Основы экономической теории / Н.Я. Кажуро. – Минск: ФАУинформ, 2001. – 672 с.

10. Экономическая теория : учебное пособие / В.Л. Клюня [и др.] ; под общ. ред. В.Л. Клюни, И.В. Новиковой. – Минск : ТетраСистемс, 2001. – 400 с.

11. Курс экономической теории : учебное пособие ; под общ. ред. М.Н. Чепурина, Е.А. Киселевой. – Киров, 1994. – 624 с.

12. Экономическая теория : учебник / В.И. Антипина [и др.] ; под общ. ред. И.П. Николаевой. – 2-е изд., перераб. и доп. – М. : ТК Велби ; Проспект, 2002. – 576 с.

13. Косова, М.М. Описательная статистика учебных текстов по физике / М.М. Косова, М.А. Зильберглейт // Труды БГТУ. Сер. VI. Физ.-мат. науки и информатика. – 2006. – Вып. XIV. – С. 167–170.

14. Невдах, М.М. Новая классификация методов определения понимания текста / М.М. Невдах, Ю.Ф. Шпаковский // Труды Белорусск. гос. технолог. ун-та. Сер. IX. Издательское дело и полиграфия. – 2007. – Вып. XV. – С. 100–104.

15. Количественные методы в исторических исследованиях ; под ред. И.Д. Ковальченко. – М. : Высшая школа, 1984. – 384 с.

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

17. Lorge, I. Predicting readability / I. Lorge // Teacher's College Record. – 1944. – № 45. – P. 404–419.

18. Микк, Я.А. Оптимизация сложности учебного текста: в помощь авторам и редакторам / Я.А. Микк. – М. : Просвещение, 1981. – 119 с.

Поступила 24.03.11

Белорусский государственный
технологический университет,
Минск, Свердлова, 13а
e-mail: yury_s@tut.by

M.A. Zilbergleit, M.M. Neudakh, Y.F. Shpakovsky

ESTIMATION OF DIFFICULTY OF UNDERSTANDING OF EDUCATIONAL TEXTS FOR HIGHER EDUCATION

Experiments with the use of various techniques such as a technique of addition, methods of expert estimations and pair wise comparisons for reception of objective criteria of the difficulty of texts are performed. As many as 49 quantitative features of educational texts were suggested and calculated. Feature space is reduced based on multidimensional statistical analysis methods. A decision rule for estimating the difficulty of understanding of teaching materials for higher education is formulated using discriminate analysis.