РАЗРАБОТКА МЕТОДА АВТОМАТИЗИРОВАННОЙ ОЦЕНКИ СЛОЖНОСТИ УЧЕБНЫХ ТЕКСТОВ ДЛЯ ВЫСШЕЙ ШКОЛЫ

М. М. Невдах

Белорусский государственный технологический университет
Минск, Республика Беларусь
E-mail: nevdax@tut.by

В статье рассмотрены основные этапы разработки метода автоматизированной оценки сложности учебных текстов. С помощью дискриминантного анализа выделены основные признаки, влияющие на усвоение учебного текста (средняя длина абзаца в словах, средняя длина абзаца в буквах, процент слов длиной в 11 букв и больше, про-цент слов длиной в 13 букв и больше), и вычислены дискриминантные функции, на основе которых появляется возможность отнести каждый объект (текст), в том числе и неизвестный, к одной из известных групп (легкий—трудный). Полученные расчеты использованы для создания программного обеспечения, автоматизирующего оценку понятности учебного материала для высшей школы.

Ключевые слова: формула читабельности, сложность текста, дискриминантный анализ, подготовленность читателей, понятность текста, характеристики текста.

С развитием и возникновением ряда наук, в которых центральным объектом анализа выступает текст, подтвердилось предположение о том, что текст представляет собой структуру, элементы которой подчиняются законам, определяющих статистическую упорядоченность и строгую организацию. Точный характер проявляющихся закономерностей, регулярностей в языке в целом крайне сложно уловить без применения математических методов и ЭВМ. Поэтому интерес к квантитативным методам как инструменту научного и практического познания статистических свойств языковых структур повышается и обусловлен объективной реальностью.

Текст как статистическая совокупность может быть охарактеризован через множество количественных переменных, на основе которых текст как последовательность символов преобразуется в набор чисел. Особенностью этих переменных является то, что они по определению не отражают глубинных, сущностных сторон текста, они описывают только внешнюю, поверхностную сторону текста. При этом следует отметить, что многие исследователи полагают [1–6], что формальные признаки каким-то опосредованным, вероятностным образом связаны с содержательной сущностью текста. В связи с этим набор количественных признаков часто является диагностическим при решении конкретной задачи (например, атрибуции текста, оценке его трудности), что, несомненно, открывает путь для проникновения в глубинную организацию текста, не доступную непосредственному наблюдению.

Квантитативный анализ текстов в настоящее время позволяет решать различные задачи, например, оценивать близость и однородность стилей текстов, классифицировать их по различным параметрам. В настоящее время в редакционно-издательской деятельности возник ряд вопросов, связанных с систематизацией и изучением текстов. Ре-

шение данных вопросов позволит, во-первых, тестировать стили авторского коллектива (в случае, когда несколько авторов пишут одну книгу) на предмет их близости, однородности. Это особенно важно в сфере учебного книгоиздания. Во-вторых, можно проанализировать лингвостатистические характеристики текстов и дать рекомендации по их корректировке. И, в-третьих, можно установить атрибуцию текста, что очень важно для текстологической науки.

В отечественной науке в настоящее время практически отсутствуют объективные инструменты для классификации текстов в зависимости от подготовленности читателей. В определенной степени вопросы количественного анализа текстов и выявления факторов, влияющих на усвоение материала, раскрыты в работах, связанных с читабельностью текста. В самом широком смысле под читабельностью понимают 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 группы понимает исследуемый материал, в какой мере скорость, с которой он его читает, приближается к оптимальной, и какой интерес представляет данный материал для читателя [7].

Первую формулу читабельности разработали В. Лайвли и С. Пресси в США в 1923 г. [8]. Позднее было создано множество формул, в которых использовались те или иные факторы трудности текста. Наибольшее распространение в США получила формула Р. Флеша [9], в котором использовалось всего два параметра: средняя длина предложения в словах и средняя длина слова в слогах.

В отечественной науке известны формулы читабельности М. С. Мацковского [10], Я. А. Микка [11] и Ю. Тулдавы [12]. В 80-х годах прошлого столетия были разработаны компьютерные программы для оценки читабельности текста: Readability Calculations, Intext, Nisus Writer и др. Разработанные продукты предназначены для анализа английского, немецкого и других языков (но не русского).

Таким образом, в настоящее время отсутствуют исследования в области читабельности с использованием современных информационных технологий и необходимого инструментария для классификации русскоязычных текстов по ряду областей знаний в зависимости от подготовленности читателя

В связи с этим был решен ряд задач. Экспериментальным материалом послужили учебные издания для вузов по философии и экономической теории. Всего было отобрано 32 отрывка длиной 1800–2000 печатных знаков. Выбор данной величины обусловлен тем, что в [13] показано, что, начиная с объема в 1800 печатных знаков, статистические характеристики текста становятся относительно постоянными. В основном эксперименте приняли участие 75 студентов Белорусского государственного технологического университета.

На первом этапе для оценки трудности понимания учебного материала использовались следующие методы: методика дополнения, экспертные оценки трудности текста и метод парных сравнений. Обработка и анализ результатов экспериментов позволили выявить необходимую информацию относительно трудности понимания учебного материала.

На втором этапе были выявлены текстовые параметры, величины которых позволили оценить сложность текста. Всего было выделено 49 признаков: 1) длина текста в абзацах; 2) длина текста в словах; 3) длина текста в буквах; 4) средняя длина абзаца в 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 са-

мостоятельного 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.

Использование такого большого числа характеристик для практических целей вызывает определенные трудности. В первую очередь это связано с тем, что данные параметры могут быть сильно коррелированны. С другой стороны, ничем не оправданное уменьшение числа переменных может привести к потере точности экспериментов. Для снижения признакового пространства были использованы следующие методы: кластерный и факторный анализы, метод корреляционных плеяд и вроцлавской таксономии, многомерное шкалирование.

Кластерный анализ представляет собой многомерную статистическую процедуру, выполняющую сбор данных, содержащих информацию о выборке объектов и затем упорядочивающую объекты в сравнительно однородные группы. 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 информационных характеристик текста использовались следующие основные алгоритмы метода кластерного анализа: межгрупповое связывание, внутригрупповое связывание, одиночное связывание, полное связывание, центроидная кластеризация, центральное связывание, метод Варда. Количество кластеров по каждому алгоритму варьировалось от 3 до 10. После выбора всех соответствующих параметров была получена необходимая информация по формированию кластеров: порядок объединения кластеров, расстояние между ними, а также принадлежность характеристик текста к тому или иному кластеру [14].

Снижение размерности набора переменных в методах факторного анализа базируется в основном на взаимной коррелированности исходных признаков. В связи с этим была вычислена корреляционная матрица, а затем выделены факторы, объясняющие разброс дисперсии.

Метод корреляционных плеяд предназначен для нахождения таких групп признаков (плеяд), в которых корреляционная связь между параметрами одной группы (внутриплеядная связь) велика, а связь между параметрами из разных групп (межплеядная связь) — мала. По определенному правилу по корреляционной матрице признаков образуют граф, который затем с помощью различных приемов разбивают на подграфы. Элементы, соответствующие каждому из подграфов, и образуют плеяду [15].

С помощью метода вроцлавской таксономии было получено нелинейное упорядочение изучаемых единиц, называемое дендритом [16]. При этом необходимо, чтобы смежные единицы дендрита имели наименее различающиеся значения признаков.

Основным преимуществом многомерного шкалирования является возможность очень наглядного визуального сравнения объектов анализа. Данный метод имеет много общего с факторным анализом, так как в обоих случаях создается система координат пространства, в котором определяется расположение точек. Однако в отличие от факторного анализа для снижения размерности используются не коэффициенты корреляции, а меры различия между объектами (расстояние Евклида; квадрат расстояния Евклида; косинус угла; неравенство Чебышева; расстояние Минковского; манхэттенское расстояние).

Основная задача многомерного шкалирования заключалась в преобразовании исходной матрицы 49×49 в гораздо более простую матрицу 49×2 и визуальным представлением ее в виде диаграммы.

После проведения перечисленных методов многомерного статистического анализа была получена необходимая информация относительно взаимосвязи признаков. Каждый из методов позволил выделить условные группы признаков, исходя из их близости.

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, подробно изученная С. Кульбаком [17]. 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 $P(x_{ij}/A_2)$ — вероятность 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 мы сократили число признаков до возможного минимума. Но этого еще недостаточно. В работах И. Лорджа [18] и Р. Флеша [9] 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.

Для дальнейшего исследования характеристик текста и их влияния на понятность учебного материала использовался дискриминантный анализ, который на основании некоторых признаков (в нашем случае характеристик текста) позволяет предсказать принадлежность объектов к двум или более непересекающимся группам. В данном случае ответы испытуемых, полученные на первом этапе, были разделены на две группы (трудный — легкий текст для восприятия). Основанием для разделения на две группы была средняя величина всех ответов испытуемых (например, по тесту № 1), которая сравнивалась со значением середины диапазона всех полученных ответов. Если среднее значение превышало середину диапазона, то разумно предположить, что текст легкий, и наоборот.

После проведения дискриминантного анализа по всем экспериментальным методикам были получены следующие функции:

$$Y = -16{,}7837 + 0{,}7602X_5 - 0{,}1002X_6 + 1{,}4484X_{32} + 0{,}0283X_{34}.$$

$$Y = -20{,}3376 + 0{,}4448X_5 - 0{,}0419X_6 + 1{,}0521X_{32} + 0{,}6791X_{34}.$$

Точность классификации при данном наборе дискриминантных переменных составляет 93,75% (30 из 32 правильных предсказаний в отношении известных объектов).

Таким образом, дискриминантный анализ позволил выявить следующие основные факторы трудности учебного текста: средняя длина абзаца в словах; средняя длина абзаца в буквах; процент слов длиной в 11 букв и больше; процент слов длиной в 13 букв и больше. На основе полученных функций было создано программное обеспечение для автоматизированной оценки читабельности учебного материала для будущих читателей.

ЛИТЕРАТУРА

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

2. *Chall, J. S.* Readability: an appraisal of research and application / J. S. Chall. Columbus, OH: Ohio State University Press, 1958.

3. *Clark, H. H.* Comprehension and the given-new contract / H. H. Clark, S. E. Haviland. Norwood NJ: Ablex Press, 1977. P. 1–40.

4. *Coleman, E. B.* Learning of prose written in four grammatical transformations / E. B. Coleman // Journal of applied psychology. 1966. № 49. P. 332–341.

5. *Dolch, E. W.* Fact burden and reading difficulty / E. W. Dolch // Elementary English review. 1939. № 16. P. 135–138.

6. *Meyer, B. J.* Reading research and the composition teacher: the importance of plans / B. J. Meyer // College composition and communication. 1982. № 1. P. 37–49.

7. *Dale, E.* The concept of readability / E. Dale, J. S. Chall // Elementary English. 1949. № 26. P. 23.

8. *Lively, B. A.* A method for measuring the vocabulary burden of textbooks / B. A. Lively, S. L. Pressey // Educational administration and supervision. 1923. № 9. P. 389–398.

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

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

11. *Микк, Я. А.* Методика разработки формул читабельности // Советская педагогика и школа. Тарту, 1974. Вып. 9. С. 78–163.

12. *Тулдава, Ю.* Об измерении трудности текстов / Ю. Тулдава // Ученые записки Тартуского университета. 1975. Вып. 345, IV, Труды по методике преподавания иностранных языков. С. 102–120.

13. *Косова, М. М.* Описательная статистика учебных текстов по физике / М. М. Косова, М.А. Зильберглейт // Труды БГТУ. Сер. VI. Издат. дело и полиграфия. 2006. Вып. XIV. С. 167–170.

14. *Невдах, М. М.* Применение кластерного анализа для исследования информационных характеристик текста / М. А. Зильберглейт, М. М. Невдах // Электроника инфо. 2008. № 1. С. 39–42.

15. *Невдах, М. М.* Анализ информационных характеристик учебных текстов с использованием эвристического метода корреляционных плеяд / М. М. Невдах // Электроника инфо. 2008. № 5. С. 47—50.

16. *Невдах, М. М.* Упорядочение характеристик текста по философии методом вроцлавской таксономии / М. М. Невдах // Сборник докладов международной научно-практической конференции студентов, аспирантов и молодых ученых. Губкин: ИП Уваров В. М., 2008. Часть 1. С. 165–168.

17. *Кульбак, С.* Теория информации и статистика / С. Кульбак. М., 1967.

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