Проверка сложности выпускных работ учащихся и студентов на основе статистических параметров

Л.В. Устинова, А.Н. Адекенова, О.В. Литвинова.

В XX веке появился ряд дисциплин прикладного характера на стыке лингвистики, математики и информатики. В частности, статистическая лингвистика — это дисциплина, изучающая количественные закономерности естественного языка, проявляющиеся в текстах. В ее основе лежит предположение, что некоторые численные характеристики и функциональные зависимости между ними, полученные для ограниченной совокупности текстов, характеризуют язык в целом или его функциональные стили (публицистический, художественный, научный и т. п.). Накопленные данные используются для дешифровки исторических письменностей, для решения задач стенографии, теории связи, информатики, а также выявления особенностей стиля отдельных авторов и атрибуции текстов. На данный момент существует ряд исследований, в которых предложены математические модели оценки сложности текста и учебных текстов с учетом возрастных особенностей учащихся. Однако, эти модели получены в основном для английских текстов, и не подкреплены соответствующими системами автоматизированного анализа. Между тем, необходимость подобных систем и соответствующих методик анализа текстов возникает у экспертов-методистов, создателей учебников, а также учителей, разрабатывающих различные методические материалы. С развитием системы экспертизы, сертификации учебной и методической литературы эти системы нуждаются в объективных и быстро реализуемых оценках ряда параметров сложности учебных текстов [1]. Задачей нашего исследования является изучение количественной оценки сложности текста. В качестве основных критериев используются статистические параметры текста, такие, как длина слова, средняя длина предложения, процент многосложных слов и другие.

Устинова, Л.В., Адекенова, А.Н., Литвинова, О.В. Проверка сложности выпускных работ учащихся и студентов на основе статистических параметров / Молодой ученый. — 2015. — № 8 (88). — С. 148-152. — URL: https://moluch.ru/archive/88/16986/ (дата обращения: 23.09.2020).

Названные параметры требуют достаточно сложных методов и технологий определения. Полученные на основе этих параметров различные формулы оценивают так называемую удобочитаемость или сложность текста. Эти параметры легко поддаются количественному выражению и пригодны для автоматической оценки. Следует отметить, что формулы удобочитаемости не являются единственным критерием качества восприятия текста, они не оценивают тонкостей авторского стиля, но чётко отличают ясный простой текст от сложного. Целью работы является разработка макропакета для автоматизации оценки сложности учебных текстов. Программа может применяться в учебном процессе для верификации курсовых и дипломных работ и определения соответствия публикаций стилю научной статьи. Автоматический классификатор функционального стиля текста создан на базе текстов, относящихся к четырём различным функциональным стилям. Критерием классификации является спектр длин слов. В ходе создания макропакета были проанализированы существующие программы поиска и анализа текстовой информации: продукт Кирсанова компании «Гарант-Парк-Интернет», инструмент удобочитаемости, «Худломер», «Орфограммка». Данные программы используются для решения следующих задач: анализ и классификация текстов, автоматическое реферирование; различные варианты поиска текста; морфологический, синтаксический и семантический анализ текста; средства навигации по большим массивам текстов. Принцип работы макропакета основывается на следующих оценках: формула Флеша (Flesch readability formula); формула Флеша-Кинкейда (Flesch-Kincaid Grade Level); индекс туманности Ганнинга (Gunning Fog Index); график читабельности текста по Фраю (Fry Readability graph); оценка читабельности Рэйгора (Raygor Readability Estimate). Для вывода статистики удобочитаемости с учетом нескольких языков нами разработан макрос на языке Visual Basic for

Устинова, Л.В., Адекенова, А.Н., Литвинова, О.В. Проверка сложности выпускных работ учащихся и студентов на основе статистических параметров / Молодой ученый. — 2015. — № 8 (88). — С. 148-152. — URL: https://moluch.ru/archive/88/16986/ (дата обращения: 23.09.2020).

Application (VBA). Такой выбор связан с популярностью текстового редактора MSWord. Макрос анализирует текст или любой его фрагмент на русском или английском языках. Отчет представлен в виде таблицы значений статистических характеристик текста (среднее число слов в предложении, среднее число слогов в слове, число многосложных слов) и оценки его сложности, в соответствии с формулой Флеша.

Таблица 1
Результаты оценки сложности текста

Всего в тексте	**Количество**		
Слов	4099		
Символов	30865		
Абзацев	297		
Предложений в абзаце	453		
Среднее количество			
Предложений в абзаце	1,5		
Слов в предложении	8		
Символов в тексте	7		
Средняя длина слова:	6,81		
Показатели легкости чтения		**Уровень**	**Шкала оценки**
Уровень образования (FKincadeE)	11,2	уровень студентов	1–20
Легкость чтения (FlashRE)	42,55	уровень студентов	0–100
Fog Index	13,23	уровень студента 1 курса	0–20
Число сложных фраз	3,3		%
Благозвучие	88,8		0–100
Дисперсия	14,58		

Устинова, Л.В., Адекенова, А.Н., Литвинова, О.В. Проверка сложности выпускных работ учащихся и студентов на основе статистических параметров / Молодой ученый. — 2015. — № 8 (88). — С. 148-152. — URL: https://moluch.ru/archive/88/16986/ (дата обращения: 23.09.2020).

Теоретическая дисперсия	15,84		

Результатом выполнения макроса является соответствие или несоответствие стиля уровню курсовой работы, с выводом дополнительных характеристик проверяемого текста (рис. 1).

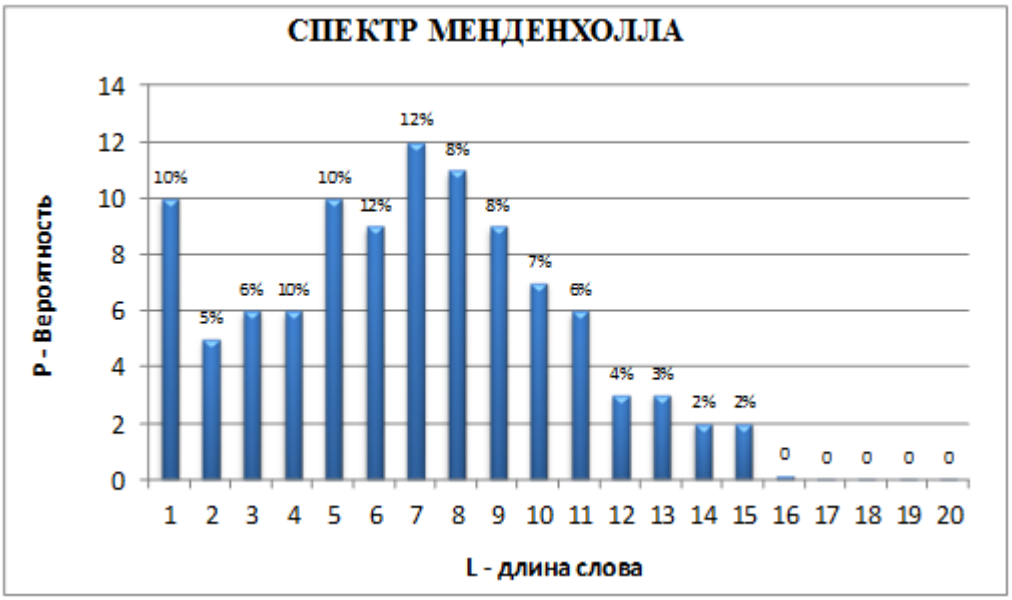


Рис. 1. Дополнительные характеристики

В основе всех, указанных выше, оценок лежит формула читаемости Флеша (Flesch readability formula), которая позволяет оценить удобочитаемость текстовых материалов. Проверка удобочитаемости по Флешу оценивается по 100-балльной шкале. Чем выше значение, тем понятнее текст. Для большинства текстов рекомендуемым значением является диапазон от 60 до 70 баллов. Формула расчета показателя удобочитаемости по Флешу:

$$ИндексФлеша_{(англ.)} = 206.835 - 1{,}015 * ASL - 84.6 * ASW,$$

где ASL — среднее число слов в предложении (число слов, деленное на число предложений); ASW — среднее число слогов в слове (число слогов, деленное на число слов). Формула Флеша является наиболее совершенной и

распространённой для определения параметров удобочитаемости текста. Флеш определил основные характеристики текста, оказывающие влияние на его восприятие. Это среднее число слогов в слове и средняя длина предложения. Свои выводы он сделал на основе исследования текстов «Экзаменационных уроков для чтения», которые традиционно используются в американской школе при переводе учеников из одного класса в другой. Данная методика получила название «формулы читабельности Флеша». Формула читаемости Флеша, скорректированная для русского языка, прогнозирует лёгкость чтения письменного материала [2]:

$$ИндексФлеша_{(рус)} = 206.835 - 1{,}3 * ASL - 60.1 * ASW$$

Именно формула соотношения этих характеристик тесно связана с уровнем понимания текста учеником.

Таблица 2
Проверка удобочитаемости по Флешу

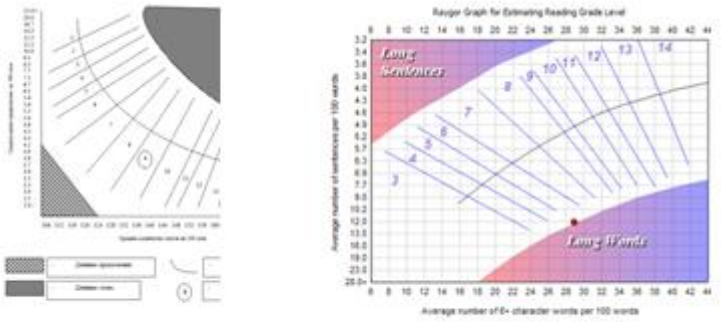
Показатель	Уровень образования
91–100	Ученик 5 класса
81–90	Ученик 6 класса
71–80	Ученик 7 класса
61–70	Ученик 8–9 классов
51–60	Выпускник средней школы
31–50	Студент института
0–30	Выпускник института

Аналогом индекса Флеша является индекс Флеша-Кинкейда. Уровень образования основан на образовательном индексе Флеша—Кинкейда и показывает, каким уровнем образования должен обладать читатель проверяемого документа. Школьный тест по Флешу-Кинкейду используется

для оценки текстов на экзаменах в школах США. Подсчет показателя делается на основе вычисления среднего числа слогов в слове и слов в предложении. Значение показателя варьируется от 0 до 20. Значения от 0 до 10 означают число классов школы. Следующие пять значений — от 11 до 15 соответствуют курсам высшего учебного заведения. Высшие пять значений относятся к сложным научным текстам. Формула для школьного теста Флеша-Кинкейда:

$$ИндексФлеша-Кинкейда = 0.39 * ASL + 11.8 * ASW - 15.59 .$$

Следует отметить недостаток обеих формул: линейная зависимость оценки от входных параметров. Хотя графики читабельности Фрая и Рэйгора [3] однозначно указывают на нелинейность построенной функции (рис. 2).



а) б)

Рис. 2. а) График читабельности текста по Фраю; б) График читабельности текста по Рейгору

Один из наиболее популярных методов оценки удобочитаемости текстовой информации — индекс Фога («индекс туманности»), разработанный в 1952 году американским ученым Р. Ганнингом [4] используется в американской журналистике. Он позволяет определить минимальный возраст читателя, которому будет понятен данный текст. Индекс туманности измеряет сложность чтения, исходя из средней длины

предложения и процента слов, состоящих из трех и более слогов. Чем выше индекс туманности, тем сложнее читать текст. Для оценки выбирается не меньше двух произвольных фрагмента текста, содержащих приблизительно по 100 слов. Учитывается средняя длина предложения (в словах) и среднее количество слогов в словах. Индекс туманности определяется по формуле

$$Индекс\,Ганнинга_{(англ.)} = 0.4 * (w + l),$$

где w — средняя длина предложения; l — среднее количество «длинных» слов (из трех и более слогов). Для текстов, написанных на русском языке, индекс Ганнинга вычисляется по формуле:

$$Индекс\,Ганнинга_{(рус.)} = 0.4 * \left[0.78 * \left(\frac{слов}{предложений}\right) + 100 * \left(\frac{число\,сложных\,слов}{число\,слов}\right)\right],$$

где числосложных слов — количество слов, с числом слогов больше четырёх; 0.78 — поправочный коэффициент для русского языка. Скорректированный индекс Ганнинга, показывает, какой образовательный уровень нужен для усвоения данного материала. Чем индекс меньше, тем большей аудитории он будет понятен. Значение 16–20 подходит для людей с высшим образованием, 9–10 — восьмиклассник и газетный уровень, 7–8 — школьный уровень. Вышеприведенные примеры показывают разнообразие подходов к оценке уровня удобочитаемости текстов. В созданном макропакете для определения средней длины слова используются два критерия: количество слогов или количество символов в слове. Количество символов в слове автоматически определяется средствами VBA. Для определения количества слогов в слове был использован алгоритм Ляна-Кнута, который по словарю с расставленными переносами строит компактный набор правил, позволяющий в точности эти места переносов восстановить. Преимуществом использования первого критерия является высокая скорость выполнения. Хотя использование второго критерия

перегружает память, он позволяет более точно определить оценку сложности текста. Для автоматизации проверки большого количества работ файлы (в формате doc) загружаются в одну папку, и запускается процесс проверки файлов. Результат проверки представлен в следующей таблице (3).

Таблица 3

Проверка удобочитаемости по Флешу

Имя документа	Индексы Флеша	Количество ошибок: орфJграф., синтакс.	Уровень студента	Стиль
Kurs_3Dmod(Шурыгина).doc	32	2; 5	4 курс	Курсовая работа
Stat_Ust_Fazilova.doc	30	0; 0	Выпускник	Научный стиль
Kurs_алг(Nemo).doc	40,97	5; 4	2 курс	Курсовая работа
K_kompmod(Зимина).doc	32	3; 4	4 курс (89.05)	Курсовая работа
Дата проверки				11/16/2014

Разработанный макропакет требует дополнительного тестирования для статистической обработки больших массивов текстовых учебников и других методических материалах. Данный макропакет был протестирован при проверки курсовых работ по информатике в школе НИШ г. Караганды. Также этот продукт полезен для ускорения проверки на наличие и определения количества ошибок в тексте; оценки сложности школьных учебных текстов с учетом возрастных особенностей учащихся. Автоматизация анализа сложности учебных текстов с применением информационных технологий на основе методов их количественной оценки, позволит увеличить эффективность обработки документов.

Литература:

Устинова, Л.В., Адекенова, А.Н., Литвинова, О.В. Проверка сложности выпускных работ учащихся и студентов на основе статистических параметров / Молодой ученый. — 2015. — № 8 (88). — С. 148-152. — URL: https://moluch.ru/archive/88/16986/ (дата обращения: 23.09.2020).

1. Попова Я. И., Шишкевич Е. В.. Стандартизация учебной литературы средней школы по критерию удобочитаемости // Севастопольский национальный университет ядерной энергии и промышленности. Научные ведомости БелГУ. Сер. Гуманитарные науки. — 2010. — № 12, вып.6. — С. 142–147.

2. Оборнева И. В. Автоматизация оценки качества восприятия текста // Вестник Московского городского педагогического университета. — Серия «Информатика и информатизация образования». — 2005. — № 2 (5) 2005. — С. 86–92.

3. http://en.wikipedia.org/wiki/Raygor_Estimate_Graph.

4. Рогушина Ю. В. Использование критериев оценки удобочитаемости текста для поиска информации соответствующей реальным потребностям пользователя // Проблеми програмування. — 2007. — № 3 — С. 76–87.