УДК 004.852

Д.С. Сугак
В.С. Озеранський

ДОСЛІДЖЕННЯ МЕТОДІВ ВИЗНАЧЕННЯ МІРИ ЧИТАБЕЛЬНОСТІ ТЕКСТУ

Вінницький національний технічний університет

Анотація

Доповідач розкриває зміст поняття «читабельність тексту», доводить необхідність дослідження цього параметру за допомогою програмних продуктів. Автор статті детально аналізує п'ять методів визначення міри читабельності тексту, вказуючи на переваги та недоліки кожного з них. Акцентує увагу на тому, що всі методи дослідження тексту спершу розроблені для англійської мови.

Ключові слова: індекс, слово, читабельність, текст, формула

Abstract

The speaker reveals the content of the concept of "readability of the text", demonstrates the need to explore this parameter using software. The author of the article analyzes in detail five methods of determining the readability of the text, pointing out the advantages and disadvantages of each. Emphasizes that all textual research methods are first developed for English.

Keywords: index, word, readability, text, formula

У сучасному світі, де розвиток техніки досяг найвищого рівня, машинний аналіз тексту може вказати на унікальність продукту, його академічну «нудоту», на здатність бути впливовим на читача. Нині постає не менш важливе питання щодо визначення читабельності тексту, оскільки при створенні певного інформаційного продукту необхідно враховувати вікові категорії читачів. Вимір читабельності тексту покликаний вказати на те, на скільки текст легкий для сприймання та розуміння. Звичайно, кожна людина ладна сама оцінити зрозумілість/незрозумілість тексту. Науковці процес ознайомлення людини із новим, ще не прочитаним текстом розклали на три етапи, тому рівень зацікавленості текстом залежить від дотримання видавцем основних правил його оформлення. Текст повинен не тільки мати гарний вигляд (відповідні шрифти, міжрядкові інтервали, колір, абзаци, підзаголовки, вирівнювання), а й бути зрозумілим читачам. Сучасна читацька аудиторія надзвичайно вибаглива, оскільки книжковий ринок пропонує досить велику кількість різноманітного матеріалу. Конкуренція видавництв спонукає до більш ґрунтовної роботи над кожним текстом, до залучення читачів різних вікових категорій.

Вимір читабельності визначається, враховуючи довжину речень, кількість складів у словах, кількість найбільш вживаних слів. Користувач або вставляє посилання на потрібну сторінку в полі URL, або копіює сам текст. Сервіс, провівши дослідження за п'ятьма формулами, виокремлює вікову категорію та визначає рівень освіти читачів, які зможуть зрозуміти текст. [1]

Найпростіший, найбільш популярний спосіб виміру читабельності був створений Рудольфом Флешем. Спочатку застосовували цей метод для англійської мови. Врахувавши довжину слів та речень, вживаність слів, оцінка складності тексту проводилася за формулою:

$$FRE = 206{,}835 - 1{,}015 * ASL - 84{,}6 * ASW \qquad (1.1)$$

де ASL – середня довжина речення в словах,

ASW – середня довжина слова у складах.

У результаті обчислень індексу Flesch–Kincaid readability tests за шкалою FRES визначається так:

1) текст легко читати, середня довжина речення – 12 або менше слів, відсутність слів 3, 4, 5-складових – 100;

2) якщо середня довжина речення від 15 до 20 слів, слова в реченнях переважно двоскладові – 60;

3) текст трохи важко читати, середня довжина речення не перевищує 25 слів, слова в цілому складаються з двох складів – 30;

4) текст дуже важко читати, речення складні, нараховують в середньому 37 слів, велика кількість слів, що мають 3, 4, 5 складів – 0. [2]

Оскільки цей метод дослідження розроблявся для англомовних текстів, то й результати стосуються відповідних читачів. Молодші школярі здатні зрозуміти текст з індексом 90-100, випускники школи зможуть осягнути прочитане з індексом 60-70, тільки люди з вищою освітою спроможні з'ясувати зміст тексту з індексом 0-30. З огляду на результати тестування у середині XX століття у кількох штатах США були прийняті законодавчі норми, за якими текст договору страхування повинен був складений таким чином, щоб його могла зрозуміти людина з середньою освітою. [2]

Automated readability index (автоматичний індекс легкості читання) визначає в американській системі освіти номер класу учнів, яким буде зрозумілий текст. Формула обчислення індексу така:

$$4{,}71 * \frac{C}{W} + 0{,}5 * \frac{W}{S} - 21{,}43 \tag{1.2}$$

де C – кількість букв і цифр;

W – кількість слів у тексті;

S – кількість речень.

Порівнюючи з іншими методами Automated readability index ґрунтується на кількості букв, а не складів, тому дуже швидко визначається за допомогою комп'ютерних програм і був розроблений саме для контролю міри читабельності в електричних друкарських машинках. [3]

Схожий до ARI, що враховує середню кількість знаків і речень на 100 слів, відомий Coleman–Liau index. Його дуже часто застосовують, якщо тексти за обсягом великі. Об'єктами дослідження є не склади, а окремі слова та речення. Формула обчислення індексу така:

$$CLI = 0{,}0588 * L - 0{,}296 * S - 15{,}8 \tag{1.3}$$

де L – середня кількість букв на 100 слів;

S – середня кількість речень на 100 слів.

Якщо Coleman–Liau index – 1, то зміст тексту повинні зрозуміти учні першого класу віком 6-7 років. Американські підлітки віком 17-18 років легко читатимуть текст з індексом – 12. [4]

На думку науковців, найпростіше вимірювання складності тексту (Simple Measure of Gobbledygook), яке оцінює роки навчання, необхідні для розуміння тексту. Аналізуючи текст за цим індексом читабельності, враховують кількість речень у тексті і кількість складних слів (більше 3 складів). Якщо таких елементів дуже багато, текст складний. Для розрахунку SMOG:

$$grade = 1{,}043 \sqrt{number\ of\ polysyllables * \frac{30}{number\ of\ sentences}} + 3{,}1291 \tag{1.4}$$

На жаль, ця формула має недолік, який полягає в тому, що текст, який аналізується, повинен нараховувати більше 30 речень. [5, 7]

Формула Dale–Chall була з часом удосконалена, оскільки спочатку орієнтувалася на список із 763 слів, які повинен розуміти середньостатистичний американський студент, бо решта слів вважалися складними для сприймання. Але 1995 року список розширився і досяг 3000 слів. Формула обчислення міри читабельності [6, 8]:

$$0{,}1579 \left(\frac{difficult\ words}{words} * 100 \right) + 0{,}0496 \left(\frac{words}{sentences} \right) \tag{1.5}$$

Отже, читабельність – це властивість тексту, яка визначається рівнем його сприймання. Оцінюється поліграфічний і лінгвістичний аспекти, тобто не тільки зовнішній вигляд тексту, але й його зміст.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. PROtext Читабельність тексту. Що це таке і як її поліпшити? (рос) [Електронний ресурс] / PROtext. – Режим доступу: https://protext.by/blog/pishi-legko/chitabelnost-teksta.-chto-eto-takoe-i-kak-ee-uluchshit/
2. Wikipedia Індекс читабельності (рос) [Електронний ресурс] / Wikipedia. – Режим доступу: https://ru.wikipedia.org/wiki/%D0%98%D0%BD%D0%B4%D0%B5%D0%BA%D1%81_%D1%83%D0%B4%D0%BE%D0%B1%D0%BE%D1%87%D0%B8%D1%82%D0%B0%D0%B5%D0%BC%D0%BE%D1%81%D1%82%D0%B8
3. Wikipedia Автоматичний індекс читабельності (рос) [Електронний ресурс] / Wikipedia. – Режим доступу: https://ru.wikipedia.org/wiki/%D0%90%D0%B2%D1%82%D0%BE%D0%BC%D0%B0%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B9_%D0%B8%D0%BD%D0%B4%D0%B5%D0%BA%D1%81_%D1%83%D0%B4%D0%BE%D0%B1%D0%BE%D1%87%D0%B8%D1%82%D0%B0%D0%B5%D0%BC%D0%BE%D1%81%D1%82%D0%B8
4. Wikipedia Індекс Колман-Ліану (рос) [Електронний ресурс] / Wikipedia. – Режим доступу: https://ru.wikipedia.org/wiki/%D0%98%D0%BD%D0%B4%D0%B5%D0%BA%D1%81_%D0%9A%D0%BE%D0%BB%D0%BC%D0%B0%D0%BD_%E2%80%94_%D0%9B%D0%B8%D0%B0%D1%83
5. Wikipedia SMOG (англ) [Електронний ресурс] / Wikipedia. – Режим доступу: https://en.wikipedia.org/wiki/SMOG
6. Wikipedia Формула читабельності Дейла-Чалла (англ) [Електронний ресурс] / Wikipedia. – Режим доступу: https://en.wikipedia.org/wiki/Dale%E2%80%93Chall_readability_formula
7. Колодний В.В. Застосування гештальт-ранжувань для виявлення переваг ОПР / В.В. Колоний, В.В. Зубко // «ІНТЕРНЕТ-ОСВІТА-НАУКА-2016» : Збірник матеріалів конференції. – Вінниця : ВНТУ, 2016. – С. 43-44.
8. Колодний В.В. Метод некритеріального структурування множини альтернатив за допомогою аналізу тернарних трирівневих ранжувань / В.В. Колодний, В.В. Зубко // «ІНТЕРНЕТ-ОСВІТА-НАУКА-2014» : Збірник матеріалів конференції. – Вінниця : ВНТУ, 2014. – С. 13-14.

Сугак Дарина Сергіївна - студент групи 2КН - 16б, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, e-mail: goodflo99@gmail.com

Науковий керівник: ***Озеранський Володимир Сергійович*** - кандидат технічних наук, старший викладач, Вінницький національний технічний університет, м. Вінниця, e-mail: ozeransky@ukr.net

Sugak Daryna - Department of Information Technologies and Computer Engineering, Vinnytsia National Technical University, Vinnytsia, email: goodflo99@gmail.com

Supervisor: **Ozeransky Volodumur** - Ph.D., senior lecturer, Vinnytsia National Technical University, Vinnytsia, e-mail: ozeransky@ukr.net