Readability of Research Writing and Text Variables in Readability Formulas

Mare-Anne Laane
Tallinn University of Technology
mareanne.laane@ttu.ee

***Abstract*- Focus is on text readability. Text accessibility and ease of comprehension is discussed. Readability formulas are reviewed with an analysis of merits and drawbacks provided by the formulas. Uses of the formulas and devices to facilitate readability are described.**

I. INTRODUCTION

Text accessibility is important both for writers and readers. There are a range of factors which make texts more or less accessible. Research has shown that factors which make the text difficult or less accessible, include poor linguistic structure, contextual structure, conceptual structure, and unclear operalization of the reader-writer relationship [1].

There are many definitions of readability. According to Klare [2], ”readability is the ease of understanding or comprehension due to style of writing”. Focus here is on writing style in contrast to factors like format, features of organization and content [3].

The definition that seems to be the most comprehensive, is the following [4]: ”In the broader sense, readability is the sum total (including interactions) of all elements with a given piece of printed material that affects the success which a group of readers have with it. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting.”

The number of readability formulas developed is massive – about 200 [2]. Studies regarding to readability formulas began in the 1950s with manual tools while today’s research is corpus-based and computerized.

The aim of this paper is to address text readability issues that guide writers to achieve acceptance by their discipline community and to review readability formulas, their merits and drawbacks as well as refer to their use in scientific publication.

II FUNCTIONS AND VARIABLES OF READABILITY

In general, there are three main functions of readability differentiated [2]:

- to indicate legibility of the printed text as well as its layout or typography
- to indicate the ease of reading due to the interest-value or the aesthetics of writing
- to indicate ease of understanding and comprehension due to the style of writing

However, today researchers look at typography and format under legibility.

In readability analysis, a multitude of variables that affect readability have been identified, grouped into four categories [5]:

- content (judged most significant)
- style (slightly less significant)
- format (third in significance)
- features of organization (least significant)

DuBay [3] has proposed the same variables with their components:

- content: propositions, organization, coherence
- style: semantic and syntactic elements
- design: typography, format, illustrations
- structure: chapters, headings, navigation

A significant finding was that of the four categories, only style – and variables related to it – could be measured statistically. Resulting from DuBay’s research, the following factors as the best readability indicators with the greatest impact were found:

average sentence length in words
percentage of ‘easy’ words
number of different ‘hard’ words
minimum syllabic sentence length number of explicit sentences
number of first, second and third-person pronouns
maximum syllabic sentence length
average sentence length in syllables percentage of monosyllables
number of sentences per paragraph
number of simple sentences
percentage of different words
percentage of polysyllables
number of prepositional phrases

Among other variables, punctuation, too, affects readability. However, there seems to be little awareness of punctuation rules. Punctuation is a readability and textuality factor and a tool for the prediction of meaning. According to Kirkman [6], punctuation is a source of ambiguity when omitted.

Following the identification of variables and grouping them into categories, intensive research directed to find the perfect formula of text readability began.

III READABILITY FORMULAS

An approach used to predicting readability is the usage of readability formulas [7]. These are mathematical equations constructed by linguists and readability researchers to help gauge the difficulty and complexity of a given piece of a text. The most frequently used formulas were created in the period from the 1930s to the 1970s and were constructed with a view to easy manual application. This is one reason why such formulas contain very few variables. However, today most readability formulas are computerized.

Readability formulas measure certain textual characteristics that are quantifiable. Such characteristics are described as ”semantic” if they concern the words used and ”syntactic” if they have to do with the length or structure of sentences. The two factors most commonly used in readability formulas are vocabulary difficulty, measured by either word difficulty or word length, and average sentence length, since a multitude of studies have proven them to be strongly associated with comprehension [4].

It is important to note that except for the surface-level features of texts, there are other variables that affect readability, like content and the reader’s abilities, but these cannot be measured mathematically and for that reason are not included in readability formulas.

Below the main readability formulas are briefly reviewed. Out of the vast collection of more than 200 readability formulas, four have been singled out herein from [7] .due to their popularity and their influence.

1 **The Flesch Reading Ease score** is given by the following equation:

$RE = 206.835 - 1.015\ ASL - 84.6\ ASW,$

where:

RE – reading ease (in a scale of 1 to 100)

ASL – average sentence length (the number of words divided by the number of sentences)

ASW – average number of syllables per word (the number of words divided by the number of sentences)

2 **The Dave-Chall formula** was designed to correct certain flaws of the Flesch Reading formula that uses two variables: average sentence length and a percentage of difficult words. The Dave-Chall formula is the following:

$Score = 0.1579\ PDW + 0.496\ ASL + 3.6365,$

where:

Score – reading grade of the reader who can answer 50 percent of the test questions on a passage

PDW – percentage of difficult words

ASL - as in the previous formula

3 **The Gunning Fog Index** is another commonly used readability measure. It became popular due to its ease of use. The formula is based on two variables: average sentence length and the number of words with more than two syllables per 100 words [3].

The Gunning Fog Index is given by the following equations:

$GL = 0.4\ (ASL + \text{hard words}),$

where:

GL – grade level

ASL – as in the previous formula

hard words – number of words with more than two syllables per 100 words

4 **The Bormuth Mean Cloze Formula** is considered to be one of the most accurate. It uses three variables: average sentence length in words, average word length in characters and the number of words on the original Dale-Chall [4] list of 3000 words, i.e. two variables for semantic and one for syntactic difficulty. This model is very complicated. The unit range is from 30 (very easy) to 100 (very hard).

Even though these formulas are useful and objective measures of text difficulty and comprehensibility, they have a number of flaws. Research in readability has pointed out several shortcomings of readability formulas. Some of the main limitations of the readability formulas referred to in literature [7] are as follows:

- They cannot measure conceptual difficulty because the content is not evaluated.
- They cannot check comprehensibility of expression.
- There is discrepancy in the results of readability formulas for the same text as the formulas use different variables.
- They assume that all readers are alike. Readability formulas make no distinctions based on reader’s characteristics.

IV NEW DIRECTIONS IN READABILITY MEASURES

To improve the accuracy of readability formulas, further studies are directed towards using more objective language-level criteria like frequency data extracted from corpora. The language level variable used by Anagnostou and Weir [7] is collocation frequency.

Some linguists maintain that collocations rather than individual words are the fundamental building blocks of a language. With collocations the whole is greater than the sum of its parts. This indicates that the meaning of a collocation cannot be derived directly from the meaning of its parts. This complex meaning in collocations constitutes a level of semantics that was not considered in earlier readability measures. Thus, Anagnostou and Weir are proposing collocational components as a factor of gauging readability.

It should be mentioned that there is no general agreement on the definition of a collocation. One of the definitions used more often is as follows [8]:

A collocation is a sequence of two or more consecutive words that has characteristics of a syntactic and semantic unit, and whose exact and unambiguous meaning cannot be derived directly from the meaning or connotation of its components.

A collocation is identified by three additional criteria [9]:

- Non-compositionality: The meaning of a collocation cannot be directly derived from the meaning of its parts.
- Non-substitutability: Components of a collocation cannot be substituted with similar ones and still hold its exact meaning.

- Non-modifiability: Collocations cannot be freely modified with additional lexical material or through grammatical transformations.

Among other criteria, association measures were used by Anagnostou and Weir [10] to decide whether any specific sequence of words qualifies as a collocation. In the readability formula proposed by them, the following variables are used: average collocation frequency, total number of collocation occurrences in sample text, number of different types of collocations in sample text, frequency of collocation type in corpus, and number of occurrences of collocation type in sample text.

They have used the British National Corpus as a reference corpus. From the corpus they extracted, using various tools and techniques, collocation frequency lists. A review of such collocation extraction tools is given in [10]. They propose that the frequency of a particular collocation in collocation frequency lists is an indicator of its semantic comprehensibility. The higher the frequency of occurrence of a collocation, the easier it will be to understand.

The new readability formula incorporating the average collocation frequency is expressed in the following form:

$RF = a \,.\, ACF + b \,.\, WDI + c \,.\, SDI + d,$

where:

RF – readability formula

ACF – average collocation frequency (semantic difficulty indicator on the level of collocations)

WDI - word difficulty indicator (commonly word length)

SDI – syntactic difficulty indicator (commonly sentence length

a, *b*, *c*, *d* -coefficients

His formula is combined with other factors that reflect the syntactic complexity features for sample texts, to provide an estimate of readability that the authors believe will prove more effective than conventional alternatives and which is grounded in a plausible theoretical approach to readability analysis.

V USES OF READABILITY MEASURE

Readability formulas have been used in the development of editing programs for desk-top computer packages.

Microsoft Word has a function for checking readability. All readability checkers do fundamentally the same thing: use algorithms based on word length, numbers of words in sentences and, sometimes, numbers of sentences in paragraphs, and then tell you how difficult the article is to read.

The eye and brain simply cannot read long words easily. Science is full of long words. The solution is to go short wherever possible: short words, short sentences and short paragraphs and cut out unnecessary words. Passives slow down reading and put a brake to readability.

If your average sentence length is 22 words, it is likely that your writing is not clear, there may be too much detail or content. According to Williams [11], ”a grammatical sentence becomes too long when a writer tacks on to one clause another that modifies it, and to that clause, yet one more”.

Most readers are aware that some texts, whatever their content ‘hang together’ better than others and are therefore easier to read. In part, this is a function of how they conform to expectations about rhetorical organization but is mainly a function how they ‘cohere’. Coherence refers to discourse relations which may or may not be signaled whereas cohesive devices are surface, textual indicators of interconnectiveness. Used well, these devices can greatly contribute to text readability.

According to research, paragraphs that violate the coherence and topicalization conventions yield longer reading times, poorer recall, and distortion of apparent theme.

Readability formulas are used to check publications in different areas to ensure that the published material is as easy to understand as possible. For instance, failure to understand documents like technical manuals may lead to serious consequences. Readability formulas are also used to evaluate the readability of scientific journals [12].

Editors and publishers use readability formulas to adapt a text to a given level of difficulty through a process of writing, rewriting and revising [4]. It means that readability formulas can function as rules for writing or rewriting a text to match a certain level of reading difficulty. However, some readability researchers [2] have advised against the use of formulas in the writing process and have advised to use them only for feedback, according to the following cycle: Write →Apply formula →Revise →Apply formula.

Hyland [13] points out that writing is an essential part of what academics do. He maintains that modern research lab devotes more energy to producing papers than discoveries and that scientists’ time is spent largely in discussing and preparing articles for publication in competition with other labs. Interest in academic writing has been accompanied by an interest in *how* academics write rather than simply *what* they write about.

From the communicative point of view, readability can either help or harm academic persuasion. Poor readability harms the power of persuasion while well-readable texts serve the purpose of convincing readers.

The higher the comprehensibility of the paper, the higher are the chances that the reader will proceed beyond the abstract and find it convincing enough to cite. According to editors, scientific literature is still abundant with lengthy, unclear prose that is likely to confuse readers. After all, authors are responsible for clarity and conciseness achievable through significant effort into their writing. There is a good saying – Easy reading is hard to write but lazy writing is hard to read.

Regarding to presentation in a research article, the first question is whether the paper is written well enough so that the technical contents may be evaluated. A paper which is incomprehensible is not publishable. An evaluation of the presentation is needed, in addition to the technical evaluation. Success of any paper is partially measured by the success of its text.

VI CONCLUSION

1 Text readability is a significant characteristic that has three main functions: legibility, ease of reading and ease of comprehension.

2 Readability formulas created can help evaluate the difficulty of texts but they only measure textual characteristics that are quantifiable. Today's interesting directions are corporate-based research using collocation frequency.

3 Readability formulas can help writers through word-processing software to calculate the readability of writing. Editors, too, are using readability formulas for paper evaluation.

4 It is the writers' responsibility to ensure that the text is readable and written in a concise and consistent style. Needlessly complex sentences misdirect readers. The more complex the idea, the more revisions are needed to clarify it.

5 Among a multitude of readability categories, the most important are content, style, format and organization.

REFERENCES

[1] G. Fulcher, "Text difficulty and accessibility: reading formulae and expert judgement," *System*, vol. 25, no. 4, pp. 497-513, 1997.

[2] G.R. Klare, *The Measurement of Readability*. Ames, Iowa: Iowa State University Press, 1963.

[3] W.H. DuBay, *The Principles of Readability*. Costa Mesa, CA. Impact Information, 2004.

[4] E. Dale and J.S. Chall, *Readability Revisited: The New Dale-Chall Readability Formula*. Massachussets: Brookline Books, 1995.

[5] W.S. Gray and B. Leary, *What Makes a Book Readable*. Chicago: Chicago University Press, 1935.

[6] J. Kirkman, *Good Style. Writing for Science and Technology*. London: E&F.N. Spon, 1996.

[7] N.K. Anagnostou and G.R.S. Weir, "From corpus-based collocation frequencies to readability measure," in G.R.S. Weir and T. Ozaso Eds. *Texts, Textbooks and Readability*, University of Strathclyde Publishing, Glasgow, 2007a.

[8] Y. Choueka, "Looking for needles in a haystack," *Proc. RIAO'88*, 1988, pp. 609-623.

[9] C.D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.

[10] N.K. Anagnostou and N.K. Weir, "Review of software collocations for deriving collocations," in G.R.S. Weir and T. Ozaso Eds. *Texts, Textbooks and Readability*. University of Strathclyde Publishing, Glasgow, 2007b, pp. 63-72.

[11] J.M. Williams, *Ten Lessons in Clarity and Grace*. University of Chicago, 3rd ed., 1989.

[12] C. Tekfi, "Readability Formulas: An Overview," *Journal of Documentation*, 43(3), 1987, pp. 261-273.

[13] K. Hyland and F. Salager-Meyer, *Scientific Writing. Annual Review of Information Science and Technology*, vol. 42, 2008, pp.297-338.