



DETERMINING QUALITY LEVELS FOR IMPROVING MAINTENANCE PROCESSES

GARAIS, G[abriel] E[ugen] & ENACEANU, A[lexandru] S[erban]

Abstract: *The emergence known by online press companies requires written filtered information for a better understanding and speed the understanding of texts and messages that are posted. Testing the readability of text in an online environment is important in the maintenance and optimization process for managing and offering quality content.*

Key words: *readability, maintenance, web content, text optimization*

1. INTRODUCTION

Readability of text is defined as a document that can easily be read and understood. Gunning Fog, Flesch Reading Ease, Flesch-Kincaid, SMOG (simple measure of gobbledygook), Fry readability formula, Automated Readability Index (ARI), Spache readability formula, Dale-Chall readability formula, Coleman-Liau index represent algorithmic-level models that are helping the site rank in a hierarchy of degrees of readability and are useful in filtering and sorting of certain information depending on the resulting interpretation of texts (DuBay, 2007). In this article we are presenting a formula for testing readability for the Romanian language usable in quality content maintenance processes. This formula is integrated through use of middleware technologies as API integration interfaces, application servers and web servers (Botezatu, et al. 2009).

2. READABILITY INTERNATIONAL FORMULAS

The readability formulas needed to develop a Romanian formula are presented in the next paragraphs and are tested in this article on two texts which contain web content written in Romanian language, noted with Story A and Story B.

Story A - <http://www.amosnews.ro/Story-29-50027>

Story B - <http://www.amosnews.ro/Story-29-50235>

We present in table 1, common parameters which are needed at the basis of the tested readability formulas.

Measured parameter	Story A	Story B
Characters	12903	528
Letters	10466	413
Phrases	109	7
Words	2120	91
Distinct words	932	49
Average words / sentence	19.45	13
Average syllables / word	2.02	1.87
Words with ≥ 3 syllables	611	25
Total count of syllables	4275	170
Percent of words ≥ 3 syllables	28.82	27.47

Tab. 1. Analyzed parameters to calculate the readability formulas

Readability formulas are divided in two categories L_1 and L_2 which are differentiated through the way of interpreting the final result. There are results that:

- Are distributed on a 0 to 100 scale;
- Indicate the level of necessary education to understand the text.

The results of formulas that take account of number of syllables L_1 , is transposed on a 0 to 100 scale, in which 0 gives the text a lower level of readability (a hard to understand text), and 100 gives the text a high level of readability (text easy to understand).

Flesch Reading Ease Model is of L_1 category with levels from 0 to 100. As the score grows higher the document is easier to understand. Web Sites must reach a level between 60 and 70 to be understood by a number of many readers.

This calculation is based on the next elements:

- Average of sentence length;
- Average number of syllables;
- The amount of personal word used;
- The amount of personal sentences used in 100 words.

The model determines how much a person with average skills can read and understand from a written message. The results are compared with determined standards for the targeted audience considering that a readable Ad contains 14 words in a sentence, 140 syllables at 100 words, 10 personal words an 43% personal sentences.

The method represents a way of verifying the communication efficiency and it is advisable using this together with other pretested processes. The formula is:

$$FRE = 206.835 - 1.015 \left(\frac{Ncv}{Pr} \right) - 84.6 \left(\frac{Tsilab}{Ncv} \right)$$

where:

FRE: Flesch Reading Ease readability formula

Tsilab: total number of syllables

Ncv: number of words

Pr: number of sentences

The coefficients 206.835, 1.015 and 84.6 (DuBay, 2007) are multiplying coefficients chosen as a result of text tests on English language. The coefficients are a consequence of a refinement process of the amount of education degree of a person that reads and understands the English language. The coefficient of 84.6 represents the amount of importance assigned to the number of words within a text. The word processors that use this algorithm are: Microsoft Word, Google Docs, Lotus WordPro, Kword.

Story A	Story B
16.5	35.6

Tab. 2. The results after applying the Flesch Reading Ease formula on story A and B

The results after applying the Flesch Reading Ease formula on the two stories A and B, demonstrates the calibration strictly for the English language being impossible for the two stories to be on such a low level on the 0 – 100 scale. The obtained result as they are can be treated as if the persons who read these texts should at least have a PhD diploma.

The researches on readability formulas show that there are formulas for next languages: Italian, Spanish, French, Danish, Japanese (DuBay, 2007).

After some tests it has been observed that the only formulas that are near as a result to the Romanian language are the formulas for the Italian and Spanish language, as it should be reasonable because of the lexical construction similarities between these languages.

The calibration of the **Flesch Reading Ease formula for the Italian language** is of L₁ category. The formula is also known as the Franchina-Vacca formula.

$$FRE_{IT} = 217 - 1.3 N_{cvmed} - 0,6 N_{sil100}$$

where:

FRE_{IT} - FLE formula for the Italian language

N_{cvmed} - average words on sentence

N_{sil100} - syllables in 100 words

Story A	Story B
70.07	88

Tab. 3. The results of FRE_{IT} formula on stories A and B

Applying the FRE_{IT} formula on story A and B shows, as in table 3, that this formula is closer to a normal level as those in table 2. So it is proved that using formulas of languages with a closer lexical form to the Romanian language is preferable.

The amount of 0.6 is applied to the number of syllables identified in 100 words chosen successively in the analyzed text and 1.3 is the amount applied to the average number of words from the total number of sentences.

The adjustment of **Flesch Reading Ease** formula for Spanish is classified as a L₁ category. The adjusted formula is known as *Fernandez Huerta*. The Spanish label comes from the name of the scientist who adjusted the initial Flesch formula.

$$FRE_{SP} = 206.84 - (0.60 * N_{sil100}) - (1.02 * N_{cvmed})$$

where:

FRE_{SP} - FLE formula adjusted for Spanish language

N_{cvmed} – number of average words from a sentence

N_{sil100} - number of syllables at 100 words

Story A	Story B
80.06	86.9

Tab. 4. The results of applying the FRE_{SP} formula on stories A and B

The result from table 4 is another prove of small gap between the lexical form of the Romanian language and others to base a new readability formula.

The models of determining readability with educational notations are of L₂ category, which can be found in specialized literature as: Gunning-Fog, Flesch-Kincaid Grade Level, SMOG, Fry, ARI, Spache, Dale-Chall, Coleman-Liau Index (Ferris & Hedgcock, 2009).

In this article it will be applied only one model of L₂ category, the **Gunning-Fog** model that shows how many years of personal education e person needs to understand with ease a specific text. A lower number denotes a better understanding and at the other point of interval, a higher number shows a more complex text and so making it hard that such a text to be understood. In this case a number of 17 need post-university education for a text to be understood. This test was created for the English language and tests mainly the number of syllables

from a word ignoring the numerical values. Testing this formula on stories A and B gives results in table 5.

$$NIV_{edu} = 0.4 * \left(\frac{Ncv}{Pr}\right) + \left(\frac{Cts}{Ncv}\right) * 100$$

where:

NIV_{edu} – US education level

Ncv - Number of words

Cts - Number of words with more than 3 syllables

Pr - Number of sentences

Story A	Story B
18.5	12.7

Tab. 5. The Gunning – Fog formula results on stories A and B

It is suggested that the number of long words should not be more than 10 to 15 at every 100 words so that texts can be understood with an education equivalent to high school.

3. READABILITY FOR ROMANIAN LANGUAGE

After many tests on > 40.000 texts of different lengths and complexity a formula was created to calculate Romanian language readability through an empirical method based on standard L₁ and L₂ formulas. The formula that results from applying the rules in determining proportions is:

$$G_{cit} = 0.0158 * \frac{L_{txt} * NIVgr}{FREts}$$

where:

G_{cit} -readability formula for texts written in Romanian language

$FREts$ - average of L₁ relations

$NIVgr$ - average of L₂ relations

L_{txt} - text length measured in number of characters

This formula determines based on readability formulas how easier or harder other texts are. From the developers point of view they have access to a table of contents which suggests them quality and quantity values. The text supervisors use the G_{cit} indicator in an automated way through filtering and calculations of an algorithm which shows them not only final results but also the intermediate stages so that they can make better decisions about keeping or improving the quality of texts that are published on the web site. The necessity of this formula comes from maintenance processes that require better contents.

4. CONCLUSION

There is not a standard for what is a quality text, but there are target audiences and for this, using the right tool can improve the experience of that target readers. The Romanian readability formula is determined empirical and must be refined in the years to come. The next step of research contains further testing for refining the formula for a reliable public use.

5. REFERENCES

Dana R. Ferris, John Hedgcock (2009) - Teaching Readers of English: Students, Texts, and Contexts, Taylor & Francis, ISBN: 978-041-5999-64-9

William H. DuBay (2007) - Unlocking Language: The Classic Studies in Readability, BookSurge Publishing, ISBN: 978-141-966-176-1

Botezatu Cornelia, Botezatu Cezar, George Carutasu, (2009) - Software integration – necessity for integrated management systems, Annals of DAAAM for 2009 & Proceedings of the 20th International DAAAM Symposium, pp 123-124, ISSN 1726-9679

**** (2010) <http://www.utexas.edu> - Texas - Austin University, Accessed on: 2010-08-18

**** (2006) <http://www.wordscout.info/hw/smog.jsp> - Smog Calculator, Accessed on: 2010-11-20