WEB APPLICATIONS READABILITY

Eugen-Gabriel Garais [1]

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 optimization process for indexing in search engines and not only.

Keywords: readability, web, text optimization

Readability of text is defined as a document that can easily be read and understood.

As [WLDB04], *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.

The models that are treated below are implemented in www.amosnews.ro dynamic site, which is made entirely by the author. This site belongs to a news agency in Romania which issues daily news to about 90. Given the large number of visitors to the news and decided to test the readability of this site for future maintenance. The large number of aggregated data sets just over 24 hours gives us a sufficient area of study as a suggestive interpretation. The text of the test models were observed that only the following tests have a consistent minimum acceptable result for the texts in Romanian and English: Flesch-Kincaid Reading Ease, Flesch Kincaid Grade Level, Gunning Fog Score, Coleman Liau I ndex, Smog Index, Automated Readability Index.

To observe the results in comparing manner there were two texts chosen and noted as story A and story B from the pages of the web site *amosnews.ro*:

Story **A** - http://www.amosnews.ro/Stire-29-50027

Story **B** - http://www.amosnews.ro/Stire-29-50235

We refer to these texts as Story **A** and Story B.

In table 1 are presented common parameters which stand at the basis of the following models

tabel 1 – Analyzed parameters to calculate the readability formulas

[1] *PhD Candidate Faculty of Computer Science for Business Management, Romanian-American University, 1B Expozitiei Blvd., district 1, code 012101, Bucharest, Romania, e-mail:garais@g4.ro*

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 nr of syllables / word	2,02	1,87
Words with >= 3 syllables	611	25
Total count of syllables	4275	170
Percent of words with >= 3 syllables	28,82	27,47

Readability formulas are divided in two categories $\mathbf{L_1}$ and $\mathbf{L_2}$ 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.

According to [DFJH09] the results of formulas that take account of number of syllables $\mathbf{L_1}$**,** 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 $\mathrm{L_1}$ 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 according to [PHWB94] 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:

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

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 are multiplying coefficients chosen according to [DFJH09] 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 according to [WWW5]: Microsof Word, Google Docs, Lotus WordPro, Kword.

table 2 – The results after applying the Flesch Reading Ease formula on story A and B

	Story A	Story B
FRE	16,5	35,6

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 shows that there are formulas for next languages: Italian, Spanish, French, Danish, Japanese according to [WWW26] and [WWW27]. The author for this article is developing and researching a formula specific for the Romanian language which is part of his PhD theses.

After some tests it has been observed that the only formulas that are close 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 L_1 category. The formula is also known as the Franchina-Vacca formula according to[WWW18] and [WWW29].

$FRE_{IT} = 217 - 1,3\ N_{cvmed} - 0,6\ N_{sil100}$

where:

FRE_{IT} - *FRE* formula for the Italian language (Franchina-Vacca)
N_{cvmed} – number of average word on sentece
N_{sil100} - number of syllables in 100 words

table 3 – The results of FRE_{IT} formula on stories A and B

	Story A	Story B
FRE_{IT}	70,07	88

Applying the FRE_{IT} formula on story A and B results as in table 3 that this formula is closer to reality 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 chose 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 L_1 category. According to WWW19, the adjusted formula is known in this field as

Fernández 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} - FRE formula adjusted for Spanish language (Franchina-Vacca)
N_{cvmed} – number of average words from a sentence
N_{sil100} - number of syllables at 100 words

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

	Story A	Story B
FRE_{SP}	80,06	86,9

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.

Adapting the **Flesch Reading Ease formula for the French languageis of** L_1 category. The adaptef romula can be found in literature under the *Kandel - Moles* name according to [WWW20], [WWW21] and [WWW30].

$FRE_{FR} = 207 – 1.015 \left(\frac{Ncv}{Pr}\right) – 73.6 \left(\frac{Tsilab}{Ncv}\right)$

where:

FRE_{FR} - FRE Kandel-Moles formula
Tsilab - total number of syllables
Ncv - number of words
Pr - number of sentences

table 5 – The FRE_{FR} formula results on stories A and B

	Story A	Story B
FRE_{FR}	38,8	56,3

Here is to mention that in spite the fact that the French lexical forms are close to the Romanian language yet the result of the readability formula shows that numbers are much to different and that this formula cannot be useful in determining a new adaptation of the Flesch formula for the Romanian language. The coefficient of 73,6 is modified from the standard of 84,6 for adapting to the *FRE* formula of the French language.

The models of determining readability with educational notations *are of L_2 category, which can be found in specialized literature as:* **Gunning-Fog, Flesch-Kincaid Grade Level, SMOG, Graficul de lizibilitate Fry, Automated Readability Index, Spache, Dale-Chall, Coleman-Liau Index.**

In this article it will be applied only one model which according to [DGJG09], the **Gunning-Fog** model 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 needs post-university education for a text to be understood. This test was created for the English language cand 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 6.

$\boldsymbol{NIV_{edu}} = 0.4*(Ncv/Pr+((Cts/Ncv)*100))$

where:

NIV_{edu} – US education level
Ncv - Number of words
Cts - Number of words with more than 3 syllables
Pr - Number of sentences

table 6 – The Gunning – Fog formula results on stories A and B

	Story A	Story B
NIV	18,5	12,7

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. After many test of more than 40.000 texts of different lengths and complexity a formula was created to calculate Romanian language readability.

The relation which results from applying the rules in determining proportions is:

$$G_{cit} = 0{,}0158 * \frac{L_{txt} * NIVgr}{FREis}$$

where:

G_{cit} - readability formula for texts write in Romanian language
$\overline{FRE\iota s}$ - the average of **FRE** relations on a 0 to 100 scale
$\overline{NIVgr}$ - average of relations which calculates the level of education needed for text understanding
L_{txt} - text length measured in number of characters

This formula determines based on complexity formulas how much other texts are easier or harder than other. This formula can be used in many case scenarios. This formula and research was done by the author having as a case study all the stories written and published by the news publishing agency Amos News (www.amosnews.ro).

From the developers point of view they have access to a table of contents which suggests them quality and quantity values. The supervisors of texts that are being added to a web site 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. Better texts can grow the number of visitors. There is not quite a standard for what is a good text, but there are target audiences and for this, using the right tool can improve the experience of that target readers.

References:

[WLDB04] William H. DuBay - *The Principles of Readability*, 2004
[DFJH09] Dana R. Ferris, John Hedgcock - *Teaching Readers of English: Students, Texts, and Contexts*, Taylor & Francis, 2009, ISBN: 978-041-5999-64-9
[WWW26] http://www.lingv.ro/resources/scm_images/RRL-34-2006-LDinu.pdf *(Conf. dr. Liviu P. Dinu (http://fmi.unibuc.ro/ro/catedre/funinf/dinu_liviu/))*
[WWW27] http://www.utexas.edu/disability/ai/resource/readability/manual/formulas-English.html *(- Univeristatea Texas - Austin)*