analytics

MDPI

Article

Readability Indices Do Not Say It All on a Text Readability

Emilio Matricciani

Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano, 20133 Milan, Italy; emilio.matricciani@polimi.it

Abstract: We propose a universal readability index, G_U, applicable to any alphabetical language and related to cognitive psychology, the theory of communication, phonics and linguistics. This index also considers readers' short-term-memory processing capacity, here modeled by the word interval I_P, namely, the number of words between two interpunctions. Any current readability formula does not consider I_p, but scatterplots of I_p versus a readability index show that texts with the same readability index can have very different I_p, ranging from 4 to 9, practically Miller's range, which refers to 95% of readers. It is unlikely that I_P has no impact on reading difficulty. The examples shown are taken from Italian and English Literatures, and from the translations of *The New Testament* in Latin and in contemporary languages. We also propose an extremely compact formula, relating the capacity of human short-term memory to the difficulty of reading a text. It should synthetically model human reading difficulty, a kind of "footprint" of humans. However, further experimental and multidisciplinary work is necessary to confirm our conjecture about the dependence of a readability index on a reader's short-term-memory capacity.

Keywords: alphabetical languages; ARI; English literature; Flesch Reading Ease Index; GULPEASE; human footprint; Italian literature; Miller's Law; short-term capacity; universal readability index; word interval

Citation: Matricciani, E. Readability Indices Do Not Say It All on a Text Readability. *Analytics* **2023**, *2*, 296–314. https://doi.org/10.3390/analytics2020016

Academic Editor: Jong-Min Kim

Received: 2 February 2023
Revised: 2 March 2023
Accepted: 21 March 2023
Published: 30 March 2023

1. Introduction

First developed in the United States [1–9], readability formulae are applicable to any alphabetical language. They are based on the length of words and sentences, and therefore they allow the comparison of different texts automatically and objectively to assess the difficulty that readers may find in reading them. From the point of view of the writer, a readability formula allows the design of the best possible match between readers and texts. Many readability formulae have been proposed for English [6], and only some for very few languages [10].

In Reference [11] we have defined a global readability formula applicable to any alphabetical language, based on a calque of the readability formula used in Italian [12], both for providing it for languages that have none, and also for estimating, on common grounds, the readability of texts belonging to different languages/translations.

In fact, because an "absolute" readability formula—i.e., a formula that provides numerical indices related to a universal origin, such as "zero"—might not exist at all, the readability formula proposed in Reference [11] can be used to compare different texts, because what counts, in this comparison, is the difference between numerical values. In other words, differences give more insight than absolute values for the purpose of comparing texts [11].

As the title of this article claims, any current readability formula, however, does not say everything about a text readability, because it neglects the response of readers' short-term memory to the partial stimuli contained in a sentence, i.e., to how the words of a sentence are punctuated, a process described by the word interval I_P [13]. All readability formula neglect, in fact, the empirical connection between the short−term memory capacity of

readers (approximately described by Miller's 7 ± 2 law [14]) and the word interval I_P, which appears, at least empirically, justified and natural [11,13,15–17].

The purpose of this article is to propose a *universal* readability formula, applicable to any alphabetical language, which includes the effect of short-term memory capacity. We base this formula on the global readability formula defined in Reference [11], which we will modify by including the word interval I_P.

After this Introduction, Section 2 revisits the classical readability formula of Italian and its relationship with the Flesch Reading Ease Index and the Automated Readability Index, largely used in English texts; Section 3 summarizes the relationship between the word interval (number of words between two interpunctions, modeling the short-term memory capacity [13]) and the number of words per sentence; examples are drawn from Italian [13] and English literature [17]; Section 4 defines and discusses our proposal of a universal readability index; Section 5 proposes a synthetic readability index of humans, a kind of "footprint" that links human short-term memory to reading difficulty; finally Section 6 draws a conclusion and suggests future work.

2. A Readability Formula for Alphabetical Languages

The observation that differences are more important than absolute values in using readability formulae [13] justifies the development of a readability formula that can be used to compare texts, even those written in different languages [15]. For most languages, in fact, no readability formula has been defined, and only few adapt English formulae to their texts [10,18]. The proposed formula, of course, does not exclude using other readability formulae specifically devised for a language—e.g., the large choice for English—[4,6] but it allows the comparison, on the same ground, of the readability of texts written in any language and in translation.

For this purpose, we have proposed in Reference [11] to adopt, as a reference, the readability formula developed for Italian, known by the acronym GULPEASE [12]:

$$G = 89 - 10C_P + 300/P_F \tag{1}$$

In Equation (1) C_P is the number of characters per word, and P_F, is the number of words per sentence. Notice that, like all readability formulae, Equation (1) does not contain any reference to interpunctions (besides, of course, full stops, question marks and exclamation marks, which determine the length of sentences), and therefore it does not consider the parameter very likely linked to the short–term memory capacity, namely the word interval I_P [13].

G can be interpreted as a readability index by considering the number of years of school attended in Italy's school system (see Reference [12]), as shown in Figure 1. The larger G, the more readable the text for any number of school years.

The continuous lines shown in Figure 1 divide the quadrant into areas of the same performance of texts, such as "almost unintelligible", "very difficult", etc. For example, the area labelled "easy" indicates all combinations of values of G and school years required to declare a text "easy" to read. In all cases, it is shown that, as the number of school years of the reader increases, the readability index he/she can tolerate decreases.

In Reference [11] we have shown, for Italian literature, that the term $10C_P$ varies very little from text to text and across seven centuries, while the term $300/P_F$ varies very much and, in practice, determines the value of the readability index.

Equation (1) says that a text is more difficult to read if P_F is large, i.e., if sentences are long, and if C_P is large, i.e., if words are long. In other words, a text is easier to read if it contains short words and short sentences, a result that is predicted by any known readability formula and should be true, of course, in any language.

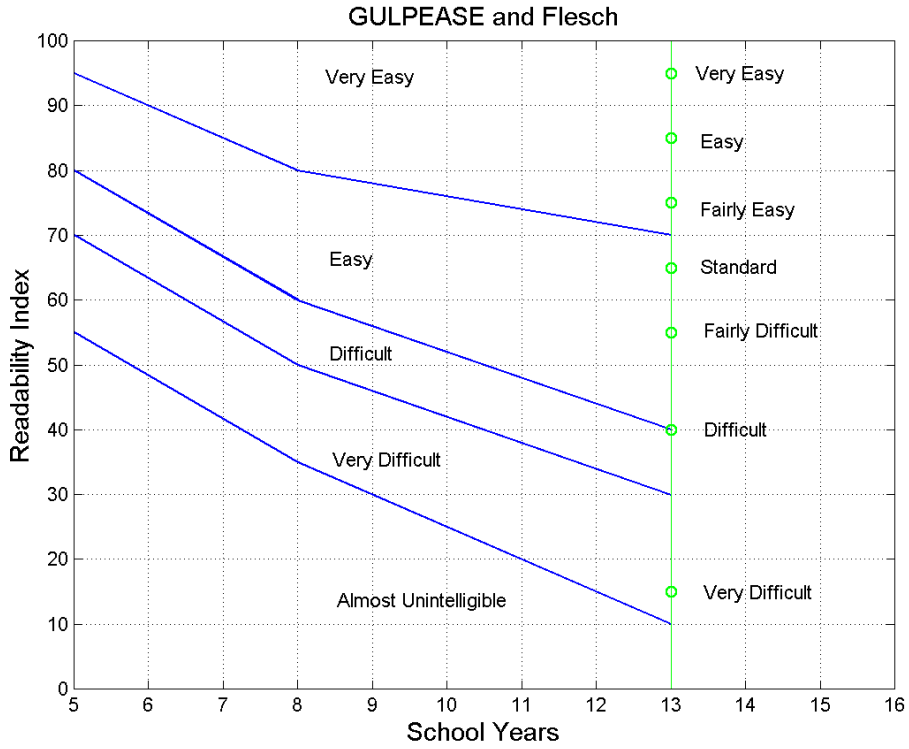


Figure 1. Readability index, *G*, of Italian (GULPEASE, see Reference [12]), as a function of the number of school years attended in Italy. The continuous lines divide the quadrant into areas of the same performance of texts. Elementary school lasts 5 years, junior high school lasts 3 years, and high school lasts 5 years. Children stay at school till they are 19 years old. For comparison, the green vertical axis on the right refers to the Flesh Reading Ease index.

In Reference [11], we have proposed the adoption of Equation (1) also for the other languages, such as those listed in Table 1, by scaling the constant 10 according to the ratio between the average number of characters per word in Italian, $< C_{p,ITA} > = 4.48$ and the average number of characters per word in another language, e.g., $< C_{p,ENG} > = 4.24$ for English. The rationale for this choice is that C_P is a parameter typical of a language which, if not scaled, would bias *G* without really quantifying the change in reading difficulty of readers, who are surely accustomed to reading, in their language, shorter or longer words, on average, than those found in Italian. This scaling, therefore, avoids changing *G* for the only reason that a language has, on average, words shorter or longer than Italian. In any case, as recalled above, C_p affects a readability formula much less than P_F [13].

Table 1. Values of C_P and k of Equations (2) and (3) in the *New Testament* texts in the indicated languages. Languages are listed according to their language family (see Reference [11]).

Language	Language Family	C_P	k
Greek	Hellenic	4.86	0.92
Latin	Italic	5.16	0.87
Esperanto	Constructed	4.43	1.01
French	Romance	4.20	1.07
Italian	Romance	4.48	1.00
Portuguese	Romance	4.43	1.01
Romanian	Romance	4.34	1.03
Spanish	Romance	4.30	1.04
Danish	Germanic	4.14	1.08
English	Germanic	4.24	1.06
Finnish	Germanic	5.90	0.76
German	Germanic	4.68	0.96

Table 1. *Cont.*

Language	Language Family	C_P	k
Icelandic	Germanic	4.34	1.03
Norwegian	Germanic	4.08	1.10
Swedish	Germanic	4.23	1.06
Bulgarian	Balto−Slavic	4.41	1.02
Czech	Balto−Slavic	4.51	0.99
Croatian	Balto−Slavic	4.39	1.02
Polish	Balto−Slavic	5.10	0.88
Russian	Balto−Slavic	4.67	0.96
Serbian	Balto−Slavic	4.24	1.06
Slovak	Balto−Slavic	4.65	0.96
Ukrainian	Balto−Slavic	4.56	0.98
Estonian	Uralic	4.89	0.92
Hungarian	Uralic	5.31	0.84
Albanian	Albanian	4.07	1.10
Armenian	Armenian	4.75	0.94
Welsh	Celtic	4.04	1.11
Basque	Isolate	6.22	0.72
Hebrew	Semitic	4.22	1.06
Cebuano	Austronesian	4.65	0.96
Tagalog	Austronesian	4.83	0.93
Chichewa	Niger−Congo	6.08	0.74
Luganda	Niger−Congo	6.23	0.72
Somali	Afro−Asiatic	5.32	0.84
Haitian	French Creole	3.37	1.33
Nahuatl	Uto−Aztecan	6.71	0.67

On the other hand, we have maintained the constant 300 because P_F depends significantly on author's style [13,15], not on language. Finally, notice that the constant 89 sets just the absolute ordinate scale, and therefore it has no impact on comparisons [13].

In conclusion, in Reference [11] we have defined a global readability index applicable to texts written in a language as:

$$G = 89 - 10kC_P + 300/P_F \tag{2}$$

with

$$k = < C_{P,ITA} > / < C_P > \tag{3}$$

By using Equations (2) and (3), we force the average value of $10 \times C_P$ of any language to be equal to that found in Italian, namely 10×4.48. Table 1 reports for Greek, Latin and 35 contemporary languages, the average values of C_P [11] and the calculated values of the constant k of Equation (3). For example, for English texts, C_P of a sample text is multiplied by 10.6, instead of 10; for Nahuatl (longer words), C_P is multiplied by 6.7, and for Haitian (shorter words) by 13.3.

Notice that k seems to be a stable factor. For example, in the sample of the English literature studied in Reference [17], we have found $< C_{P,ENG} > = 4.23$ (instead of the 4.24 of Table 1). Now, because the value found in the Italian literature [13] is $< C_{P,ITA} > = 4.67$, therefore $k = 4.67/4.23 = 1.10$, instead of the $k = 4.48/4.24 = 1.06$ of Table 1.

As recalled above, all readability formulae substantially tell the same story, and therefore they should be very similar and it is very likely that any one of them can be obtained from another. We illustrate this fact with an example.

Because English is the language that has more readability formulae than any other language, let us compare G to the most classical English readability formula proposed and amply discussed by Flesch [1,2], known as the Flesch *R*eading *E*ase (RE) formula:

$$RE = 206.8 - 1.015w - 84.6s \tag{4}$$

In Equation (4), w is the average number of words per sentence, and s is the average number of syllables per word. Because the number of characters per word is, on average, proportional to the number of syllables per word, the parameter s paralles C_P and, of course, $w = P_F$.

How Equation (4) quantifies the degree of difficulty was defined by Flesch himself [1,2], and its values are reported in the vertical scale of Figure 1 (right ordinate scale), for comparison with G (left ordinate scale). Figure 2 shows the scatterplot between the values calculated with the global readability index G, Equation (2), versus those calculated with RE, Equation (4), according to WinWord, in novels from English literature [17], Table 2.

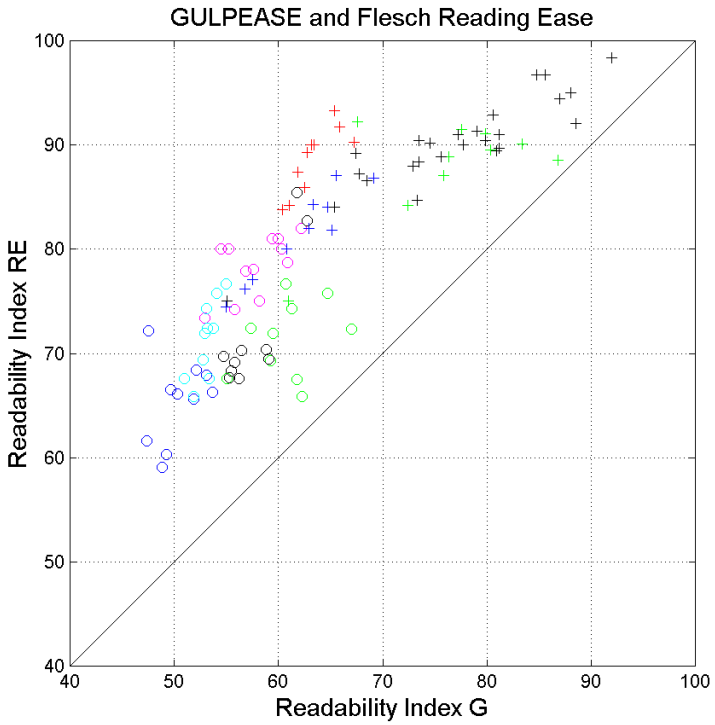


Figure 2. Flesch Reading Ease (RE) index, Equation (4), versus the global index G, Equation (2), for the novels of the English Literature listed in Table 2. *Robinson Crusoe*, cyan "o"; *Pride and Prejudice*, black "o"; *Vanity Fair*, blue "o"; *Alice's Adventures in Wonderland*, magenta "o"; *Treasure Island*, green "o"; *Adventures of Huckleberry Finn*, red "+"; *Peter Pan*, blue "+"; *The Sun Also Rises*, green "+"; *A Farewell to Arms*, black "+".

Table 2. Novels from English literature. Deep-language parameters C_P, P_F, I_P, G and universal readability index G_U, the latter discussed in Section 4. Novels are listed according to the year of publication.

Literary Work	C_p	P_F	I_P	G	G_U
Matthew King James translation (1611)	4.27	23.51	5.91	55.14	55.86
Robinson Crusoe (D. Defoe, 1719)	3.94	57.75	7.12	50.84	42.22
Pride and Prejudice (J. Austen, 1813)	4.40	24.86	7.16	52.79	43.89
Wuthering Heights (E. Brontë, 1845–1846)	4.27	25.82	5.97	53.65	53.89
Vanity Fair (W. Thackeray, 1847–1848)	4.63	25.74	6.73	49.75	44.10
David Copperfield (C. Dickens, 1849–1850)	4.04	24.40	5.61	56.68	59.66
Moby Dick (H. Melville, 1851)	4.52	31.18	6.45	49.11	45.66
The Mill on The Floss (G. Eliot, 1860)	4.29	28.03	7.09	52.70	44.32
Alice's Adventures in Wonderland (L. Carroll, 1865)	3.96	30.92	5.79	56.14	57.76

Table 2. *Cont.*

Literary Work	C_p	P_F	I_P	G	G_U
Little Women (L.M. Alcott, 1868–1869)	4.18	21.08	6.30	57.31	54.99
Treasure Island (R. L. Stevenson, 1881–1882)	4.02	21.89	6.05	58.78	58.39
Adventures of Huckleberry Finn (M. Twain, 1884)	3.85	24.89	6.63	59.01	54.14
Three Men in a Boat (J.K. Jerome, 1889)	4.25	13.71	6.14	64.19	63.13
The Picture of Dorian Gray (O. Wilde, 1890)	4.19	16.56	6.29	62.83	60.58
The Jungle Book (R. Kipling, 1894)	4.11	21.52	7.15	57.95	49.14
The War of the Worlds (H.G. Wells, 1897)	4.38	20.85	7.67	55.31	42.48
The Wonderful Wizard of Oz (L.F. Baum, 1900)	4.02	20.55	7.63	59.38	46.85
The Hound of The Baskervilles (A.C. Doyle, 1901–1902)	4.15	17.79	7.83	60.27	46.16
Peter Pan (J.M. Barrie, 1902)	4.12	18.20	6.35	60.53	57.85
A Little Princess (F.H. Burnett, 1902–1905)	4.18	16.38	6.80	61.57	55.45
Martin Eden (J. London, 1908–1909)	4.32	16.94	6.76	59.38	53.50
Women in love (D.H. Lawrence, 1920)	4.26	13.71	5.22	63.98	70.02
The Secret Adversary (A. Christie, 1922)	4.28	11.02	5.52	69.08	72.76
The Sun Also Rises (E. Hemingway, 1926)	3.92	10.70	6.02	72.58	72.45
A Farewell to Arms (H. Hemingway,1929)	3.94	10.12	6.80	73.17	66.99
Of Mice and Men (J. Steinbeck, 1937)	4.02	9.67	5.61	74.20	77.24

We can notice a fair agreement between the two indices, with a correlation coefficient of 0.850. The bias could be compensated by downscaling RE.

The attribution of the grade level GL in the USA school system was defined by Kincaid et al. [3], by using the same parameters w and s. The grade level is similar to that attributed to G.

Another readability formula, the Automated Readability Index (ARI), was also defined by Kincaid et al. ii for specific military documents [3]. It is fully related to G because it depends on the same parameters, C_P and P_F:

$$ARI = 4.71C_p + 0.5P_F - 21.43 \quad (5)$$

As ARI increases, the age of required readers increases too. Figure 3 shows the scatterplot between the global G, Equation (2), and ARI, for the the same English novels considered in Figure 2. We can see a very tight relationship for fixed C_P.

In conclusion, the global readability formula, Equation (2), provides a readability index that can be directly scaled to ARI and approximately also to RE. For this reason, we continue studying G, which we will modify by introducing the word interval I_P to obtain the universal readability formula/index mentioned above. To do so we need to recall, in the next section, some fundamental knowledge on I_P.

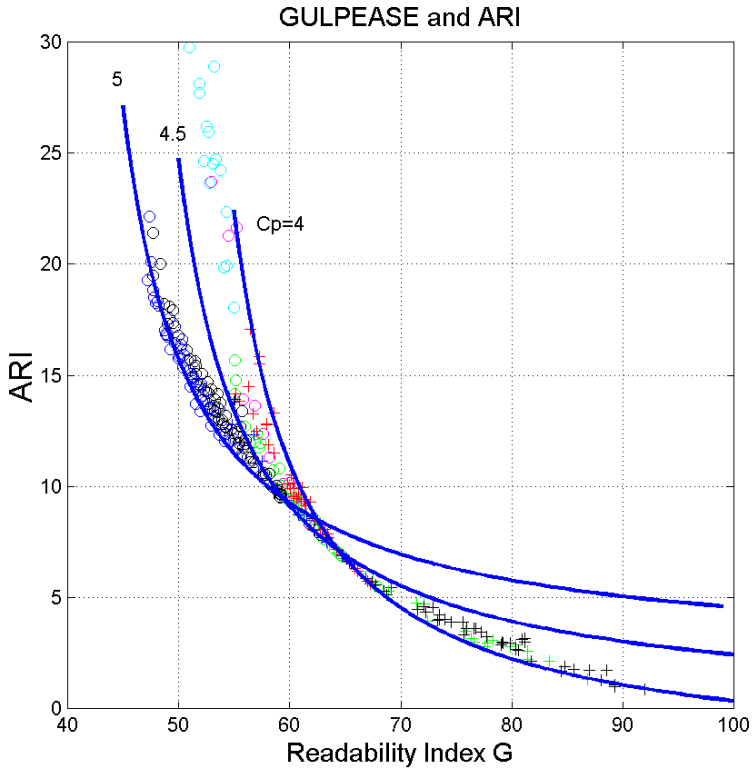


Figure 3. Automated Readability Index (ARI), Equation (5), versus the global index G, Equation (3), for the novels of English literature listed in Table 2. The continuous lines assume constant values of C_P. *Robinson Crusoe*, cyan "o"; *Pride and Prejudice*, black "o"; *Vanity Fair*, blue "o"; *Alice's Adventures in Wonderland*, magenta "o"; *Treasure Island*, green "o"; *Adventures of Huckleberry Finn*, red "+"; *Peter Pan*, blue "+"; *The Sun Also Rises*, green "+"; *A Farewell to Arms*, black "+".

3. Word Interval and Short-Term Memory

As we have discussed in References [11,13,15], the word interval I_p—namely the number of words per interpunctions—varies in the same range of the short-term memory capacity-given by Miller's 7 ± 2 law [14], a range that includes 95% of all cases, and very likely the two ranges are deeply related because interpunctions organize small portions of more complex arguments (which make a sentence) in short chunks of text, which are the natural input to short-term memory [19–27]. Moreover, I_p, drawn against the number of words per sentence, P_F, tends to approach a horizontal asymptote as P_F increases, and this occurs both in ancient classical languages (Greek and Latin) and in contemporary languages, as shown in References [11,13] by studying translations of the *New Testament* books from Greek. In other words, even if sentences get longer, I_p cannot get larger than about the upper limit of Millers' law (namely 9), because of the constraints imposed by the short-term memory capacity of readers and writers, as well.

The average value of I_p can be empirically related to the average value of P_F according to the non-linear relationship [13]:

$$< I_P >= (I_{P\infty} - 1) \times \left[1 - e^{-\frac{(<P_F>-1)}{(P_{Fo}-1)}}\right] + 1 \quad (6)$$

where $I_{P\infty}$ gives the horizontal asymptote, and P_{Fo} gives the value of $< P_F >$ at which the exponential falls at $1/e$ of its maximum value.

Equation (6) is a good average mathematical model for Italian literature [13] and also for Greek, Latin and contemporary languages [11,15]. Reference [11] reports the values of $I_{P\infty}$ and P_{Fo} for each language considered.

Presently, we have carried out the same analysis as for the large corpus of Italian literature [13] for a smaller but useful corpus of the English literature recently studied in

Reference [17], and have calculated the best-fit values of Equation (6). Figure 4 shows the scatter plot of I_p versus P_F (values calculated for each chapter) and the best-fit curve, with $I_{P\infty} = 6.70$ and $P_{Fo} = 6.78$, to be compared with $I_{P\infty} = 7.37$ and $P_{Fo} = 10.22$ of the Italian literature, whose curve is also drawn.

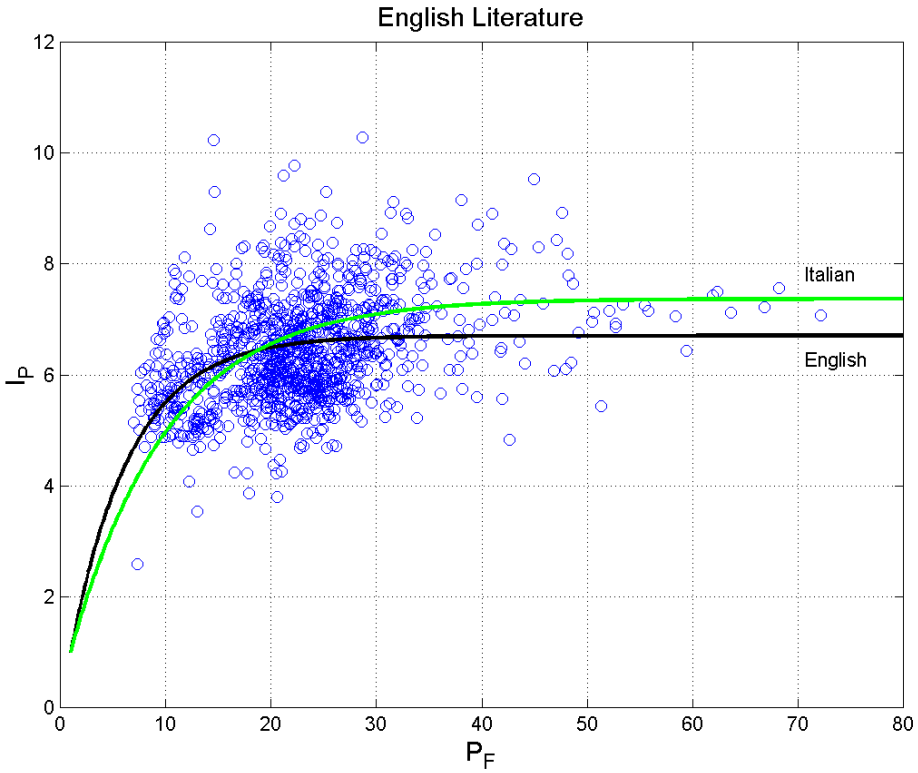
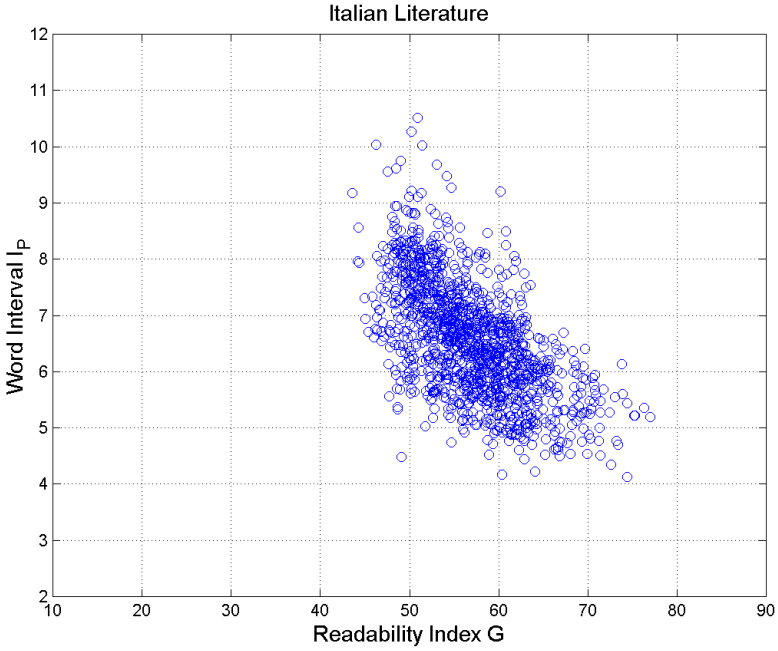


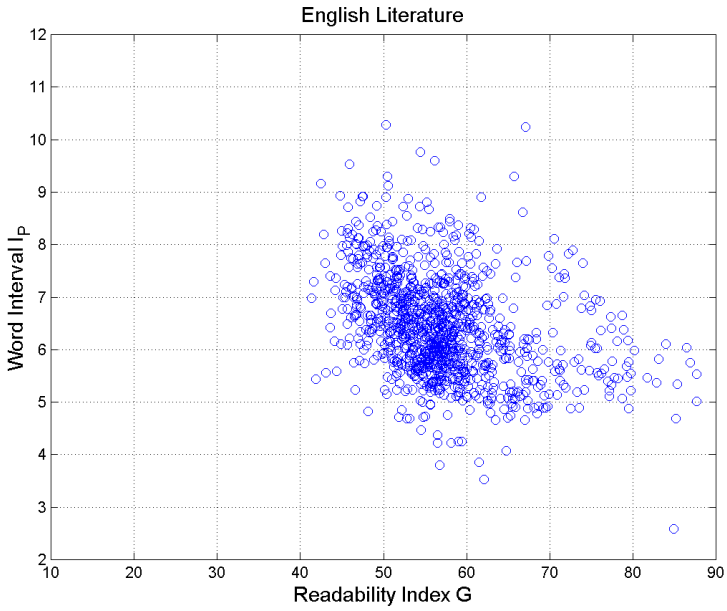
Figure 4. Scatter plot of the word interval, I_P, and the number of words per sentence, P_F, for all samples of the novels considered in the English literature, Table 2. The continuous black line refers to the best fit given by Equation (6). The green line refers to the Italian literature [13]. Miller's bounds are given by $I_P = 7 \mp 2$.

Notice that the constants of the English literature differ from those reported in Reference [17] ($I_{P\infty} = 6.57$, $P_{Fo} = 4.16$) for the same literary corpus, because the latter were the results of fitting Equation (6) to the average values of I_P and P_F, not to the values of I_P and P_F obtained by considering the samples (a sample for each chapter), which give the scatterplot drawn in Figure 4. The different values are due, of course, to the non-linear best fit.

Now, as we have recalled in Section 2, any readability index is practically a function only of P_F. Readability formulae do not consider I_p, but the scatterplots of I_p versus G show an interesting story: texts with the same G do not show the same I_p. In other words, according to the theory of readability formulae, a text with a given index should be readable with the same effort both by readers who display a powerful short-term memory processing capacity (large I_p) and by readers who do not (small I_p). For example, for $G = 60$ ("easy/standard" texts for readers with 8 years of school, Figure 1), Figure 5 shows that I_P can vary from 4 to 9. This is practically Miller's range, which refers to 95% of readers [14]. We think that these readers should be distinguished, and therefore, our aim is to propose, in the next section, a possible "universal" readability index, G_{UI}, based on G, which includes I_p.



(**a**)



(**b**)

Figure 5. Scatterplots of the readability index, G, versus the word interval, I_P. (**a**): Italian Literature [11]; (**b**): English Literature, Table 2. Miller's bounds are given by $I_P = 7 \mp 2$.

4. A Universal Readability Formula

We suppose that the global readability index G should be modified by introducing a function that depends linearly on I_P. Our hypothesis is based on Miller's law, which quantifies linearly the processing capacity of the short-term memory. Moreover, the function should not change the global value for a reader with an "average" processing short-term memory capacity. For words, this average is not 7, but about 6 [1,28]; therefore, in the following we assume this latter value. Notice that 6.03 is the average value of I_P (standard deviation 1.11) of the data listed in Table 2 of Reference [11], a further indication of its barycentric value.

We write our proposed universal readability formula as:

$$G_U = G - \frac{\Delta G}{\Delta I_P}(I_P - 6) \tag{7}$$

where G is given by Equation (2).

We assume that the numerical value of the discrete derivative $\frac{\Delta G}{\Delta I_P}$ is given by:

$$\frac{\Delta G}{\Delta I_P} = \frac{G_{max} - G_{min}}{I_{P,max} - I_{P,min}} \tag{8}$$

In Equation (8), the numerical values are the maximum and minimum averages found in the Italian literature—see Reference [13], whose oldest texts (seven centuries old, e.g., Boccaccio's *Decameron*) are still read today in Italian high schools with a reasonable effort, a possibility not available in other Western languages.

From [13], we calculate:

$$\frac{\Delta G}{\Delta I_P} = \frac{69.84 - 49.54}{8.24 - 4.94} = 6.15 \approx 6.00 \tag{9}$$

Therefore, the proposed universal readability formula is given by

$$G_U = G - 6(I_P - 6) \tag{10}$$

Equation (10) sets $G_U = G$ for $I_P = 6$; $G_U < G$ for $I_P > 6$ and $G_U > G$ for $I_P < 6$. In other words, if a text with a given G, has a small word interval I_P, then it should be read more easily than a text with the same G, but larger I_P. For example, texts with $G = 60$ would be transformed in Miller's range of 5 to 9 to $G_U = 66$ for $I_P = 5$ and in $G_U = 42$ for $I_P = 9$, and therefore, in the first case, the text considered "easy" after 8 years of school (Figure 1), is considered "easy" to read but only after 7.2 years of school; in the second case, the text would be considered "easy", but only after about 13.2 years of school. The meaningful difference between the two indices is therefore very large: $66 - 42 = 24$, corresponding to $13.2 - 7.2 = 5$ years of school. This significant difference would be lost in the original formula of Equation (2), or in any other readability formula.

Figure 6 shows the scatterplots between G_U and I_P (blue circles) for the samples concerning the literary texts considered in Italian [13] and in English Literatures, in Table 2. Compared to the scatterplots of Figure 5 (redrawn in Figure 6 with red circles), the difference between G_U and G is evident: the linear dependence of G_U on I_P, according to Equation (10), spreads the values around a line and introduces significant correlation coefficients, -0.9016 for Italian, and -0.7730 for English. The regression line:

$$G_U = -aI_P + b \tag{11}$$

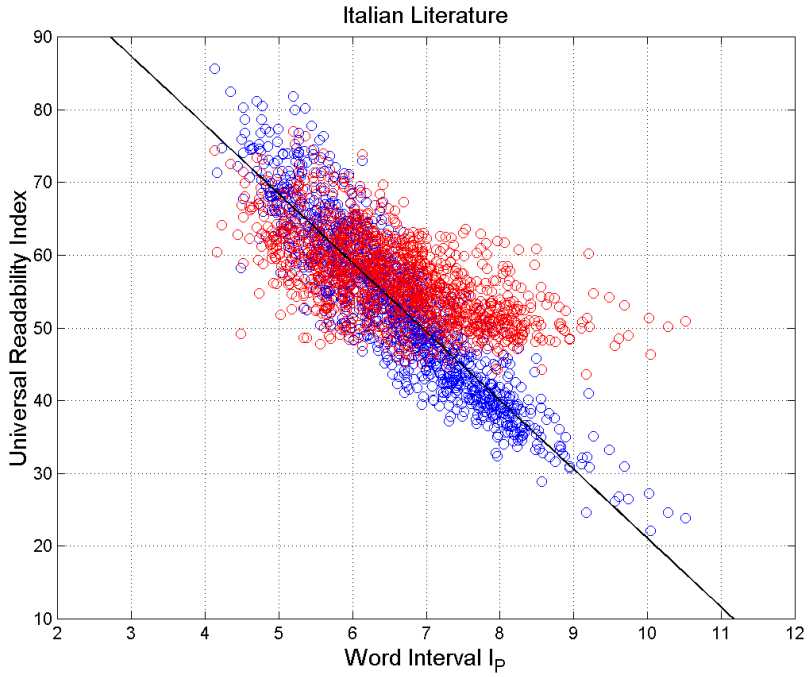
is very similar in the two languages:

$$G_{U,ITA} = -9.47I_P + 115.71 \tag{12}$$

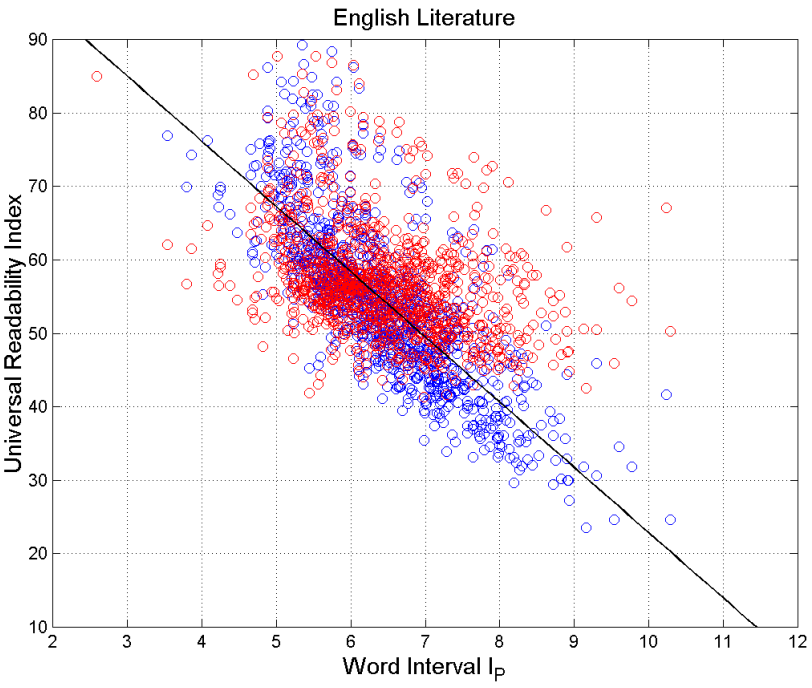
$$G_{U,ENG} = -8.88I_P + 111.64 \tag{13}$$

This result indicates that Equation (11) might be "universal".

Finally, some specific examples concerning novels taken from Italian and English literatures will further illustrate the relationship between G and G_U.



(**a**)



(**b**)

Figure 6. Scatter plot of the readability index, G_U, versus the word interval, I_P (blue circles). The red circles refer to the scatterplots of Figure 5. (**a**): Italian Literature [13]; (**b**): English Literature, Table 1. Miller's bounds are given by $I_P = 7 \mp 2$.

Table 3 shows how the readability index is modified from G to G_U for some Italian novels written from the XIV to the XX century [13]. For example, it is interesting to notice how G is transformed into G_U for the two novels written by Alessandro Manzoni.

Table 3. Novels from Italian Literature [13]. Average deep-language parameters C_P, P_F, I_P, and G and corresponding universal readability index, G_U. Novels are listed according to the alphabetical order of the author's name.

Novel	C_P	P_F	I_P	G	G_U
Anonymous (*I Fioretti di San Francesco*, XIV Century)	4.65	37.70	8.24	50.70	37.26
Boccaccio Giovanni (*Decameron*, XIV)	4.48	44.27	7.79	51.18	40.44
Buzzati Dino (*Il deserto dei tartari*, XX)	5.10	17.75	6.63	55.27	51.49
Calvino Italo (*Marcovaldo*, XX)	4.74	17.60	6.59	59.19	55.65
Cassola Carlo (*La ragazza di Bube*, XX)	4.48	11.93	5.64	69.84	72.00
Collodi Carlo (*Pinocchio*, XIX)	4.60	16.92	6.19	61.57	60.43
Deledda Grazia (*Canne al vento*, XX)	4.51	15.08	6.06	64.39	64.03
D'Annunzio Gabriele (*Le novelle delle Pescara*, XX)	4.91	17.99	6.38	58.16	55.88
Eco Umberto (*Il nome della rosa*, XX)	4.81	21.08	7.46	55.78	47.02
Fogazzaro (*Piccolo mondo antico*, XIX-XX)	4.79	16.08	6.10	61.46	60.86
Gadda (*Quer pasticciaccio brutto ...* XX)	4.76	18.43	4.98	58.24	64.36
Machiavelli Niccolò (*Il principe*, XV-XVI)	4.71	40.17	6.45	49.54	46.84
Manzoni Alessandro (*Fermo e Lucia*, XIX)	4.75	30.98	7.17	51.72	44.70
Manzoni Alessandro (*I promessi sposi*, XIX)	4.60	24.83	5.30	56.00	60.20
Moravia Alberto (*La ciociara*, XX)	4.56	29.93	7.28	53.52	45.84
Pavese Cesare (*La luna e i falò*, XX)	4.47	17.83	6.83	61.90	56.92
Pirandello Luigi (*Il fu Mattia Pascal*)	4.63	14.57	4.94	63.94	70.30
Svevo Italo (*Senilità*, XX)	4.86	16.04	7.75	59.39	48.89
Tomasi di Lampedusa (*Il gattopardo*, XX)	4.99	26.42	7.90	50.72	39.32
Verga (*I Malavoglia*, XIX-XX)	4.46	20.45	6.82	59.34	54.42

Alessandro Manzoni (Milan 1785, Milan 1873), one of the most studied Italian novelist in Italian high schools (*Licei*) and universities, in 1827 published *Fermo e Lucia* (*Fermo and Lucia*), a text that scholars of Italian Literature—and Manzoni himself—consider the "first" version of his masterpiece *I Promessi Sposi* (*The Betrothed*, available in a new English translation [29]) published in the years 1840–1842. According to scholars of Italian literature [30–33], the two versions differ very much, both in story structure and characters and, as far as we are here concerned, also in style and language; therefore, it is interesting to see how much the author transformed (mathematically) *Fermo e Lucia* into *I Promessi Sposi*, a study partially carried out in References [13,15].

As far as readability is concerned, from Table 3 we notice a large improvement in *I Promessi Sposi*, compared to *Fermo e Lucia*, if differences are considered. In fact, $G = 51.72$ in *Fermo e Lucia* and $G = 56.00$ in *I Promessi Sposi*, a difference of only 4.28 units, leading to a decrease in school years (for "easy" reading, Figure 1) of only about 0.8 years. This difference does not justify the reading difficulty of the two texts discussed by scholars of Italian literature [30–33]. However, if we consider G_U, then the difference is quite large, very likely measuring the relative reading difficulty, because G_U ranges from 44.70 to 60.20, a difference of 15.5 units leading to a decrease in school years (for "easy" reading, Figure 1) from 11.8 (*Fermo e Lucia*) to only 8 (*I Promessi Sposi*), well justified by scholars of Italian literature [30–33]. In conclusion, G_U is a better estimate than G in assessing the difference in reading difficulty between these two very studied novels.

Table 2 shows also how the readability index is modified from G to G_U in some English novels.

As we can read from Table 2, in *Robinson Crusoe* the readability index decreases from 50.84 to 42.22, therefore passing from about 10.3 to 12.4 years of school for "easy" reading

(Figure 1). For Hemingway's novels, *The Sun Also Rises* is more readable (72.45) than *A Farewell to Arms* (66.99); the order given by G, i.e., 72.58 and 73.17, respectively, is reversed, therefore reducing the number of years of school required for "easy" reading by 1 (Figure 1). *The Hound of The Baskervilles* changes its readability index from 60.27 to 46.16, therefore passing from 8 to 11.5 years of school for "easy" reading (Figure 1).

In conclusion, by introducing the word interval I_P in the definition of a readability index, as in Equation (10), readability differences in texts are more "fine-tuned" for readers.

5. A "Footprint" of Humans

As already recalled, in Reference [11] we have studied the translation of the *New Testament* from Greek to Latin and to contemporary languages. For all these translations, we have recently calculated the scatterplots between G and I_P, and between G_U and I_P, with results very similar to those shown in Figure 6. Some specific examples are reported in Appendix A. Similarly, we have calculated the linear best fit between G_U and I_P. Appendix B lists the values of the constants a and b of Equation (11) for each translation/language.

This set of values are useful because they could be used to compare texts written in any language. For example, in *David Copperfield*, G_U is estimated to be 61.82 with Equation (13) and 66.02 with the values of Appendix B (the experimental average value is 59.66, Table 2); in *The Hound of The Baskervilles*, G_U is estimated to be 42.11 with Equation (13) and 48.53 with the values of Appendix B (the experimental average value is 46.16, Table 2). In the first case, the difference in the readability of the two novels is 19.71, and in the second case it is 17.49, which implies an "error" of about 0.25 years of school (Figure 1).

It may be interesting to consider the most compact relationship between G_U and I_P, given by the overall average values of the constants reported in Appendix B:

$$G_U = -8.94 I_P + 116 \tag{14}$$

Figure 7 show this average relationship together with ± 1 standard deviaton bounds. These extremely compacted curves can synthetically represent how the capacity of human short-term memory (modelled by I_P) is related to the difficulty of reading a text, in any alphabetical language; therefore, it may be considered as a kind of "footprint" of humans.

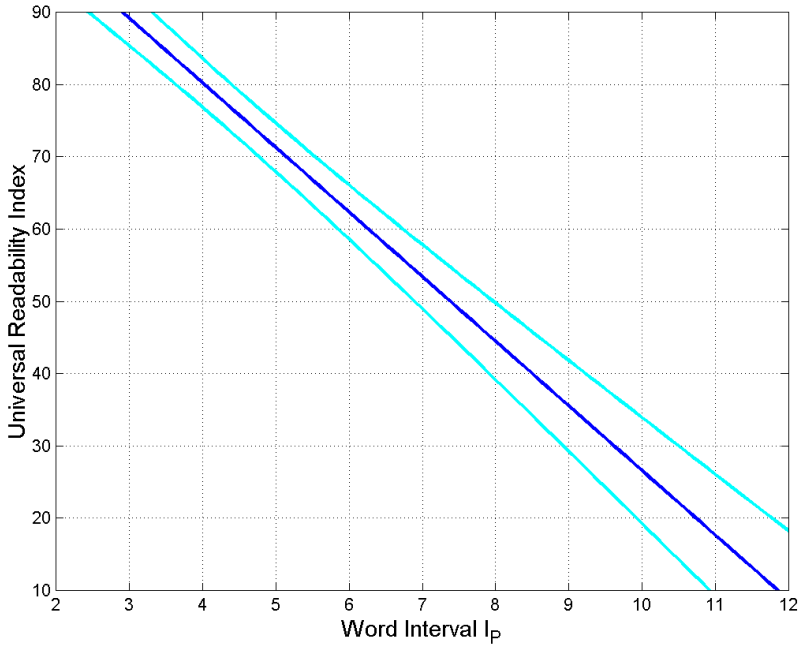


Figure 7. Average (blue line) and ± 1 standard deviation lines (cyan lines) of the universal readability index, G_U, versus the word interval, I_P, from Table A1.

6. Conclusions

We have proposed a universal readability index, G_U, Equation (10). Compared to the current readability indices, this index considers also readers' short-term memory processing capacity, here described by the word interval I_P, namely, the number of words between two interpunctions. The observation that differences give more insight than absolute values has justified, we think, the development of a universal readability formula which is useful for comparing texts written even in different languages and is applicable to alphabetical languages and related to cognitive psychology, the theory of communication, phonics and linguistics.

Scholars have never considered including in the current readability formulae the word interval, I_p, but the scatterplots of I_p versus any readability index show that texts with the same readability index can have very different values of I_p. Now, it is unlikely that I_P has no impact on reading difficulty. By introducing I_P in the definition of a readability index, readability differences in texts are better "fine-tuned" for readers, e.g., to their school years as a reference. We have used the global readability index developed for Italian [11], after showing that Flesch's index and ARI are connected to this index because they depend on the same variables.

We have calculated an extremely compact formula, Equation (14), which can measure how the capacity of human short-term memory (modelled by I_P) is likely related to the difficulty of reading a text, measured by the universal readability index G_U, here defined. We think that it synthetically models human reading difficulty, i.e., it might be considered a "footprint" of humans.

However, there is an important aspect to be considered. Because, as far as we know, there are no direct experiments on the relationship between readability and short-term memory capacity, the universal index here proposed, Equation (10), should be considered a first step in researching this important relationship. Therefore, further work needs to be carried out by a multidisciplinary team of researchers to fully validate Equation (10).

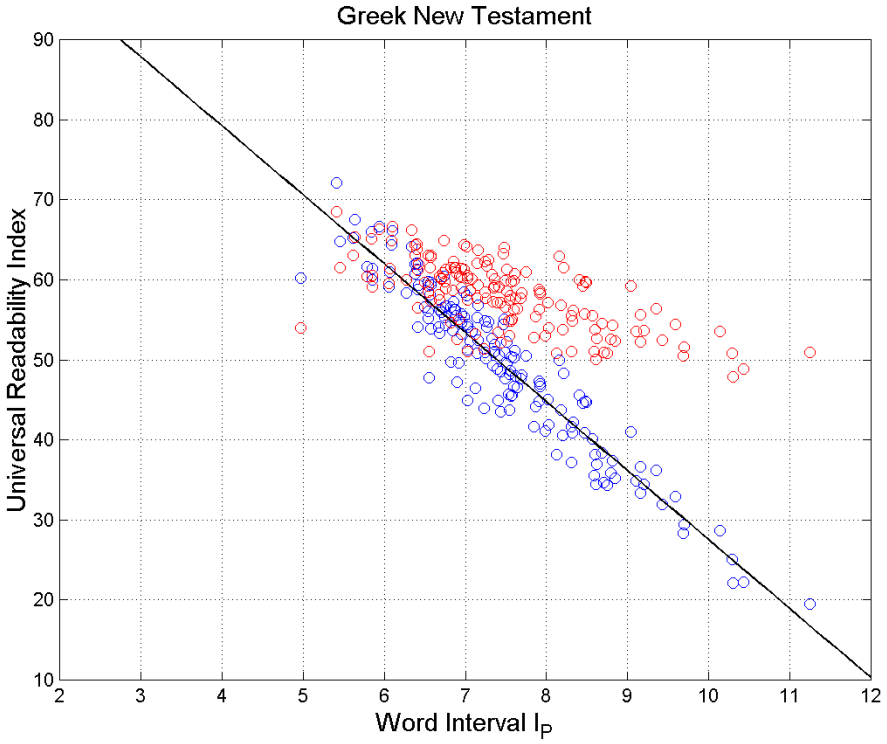
Funding: This research received no external funding.

Data Availability Statement: Data are available, on request, by the author.

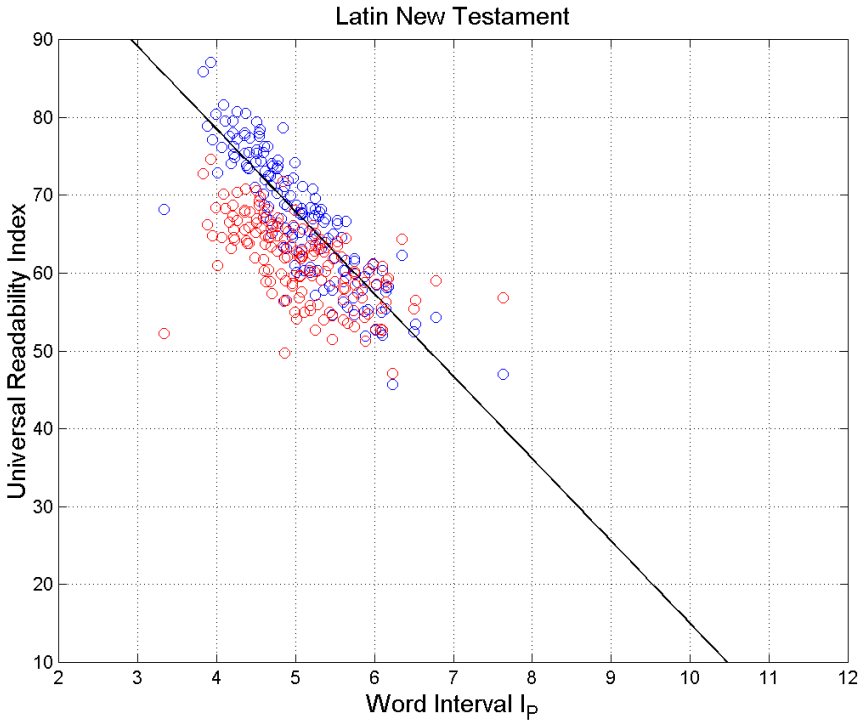
Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Scatterplots between G_U and I_P, for selected languages. We show the scatterplots between G and I_P (red circles), and between G_U and I_P, (blue circles) for some selected languages.

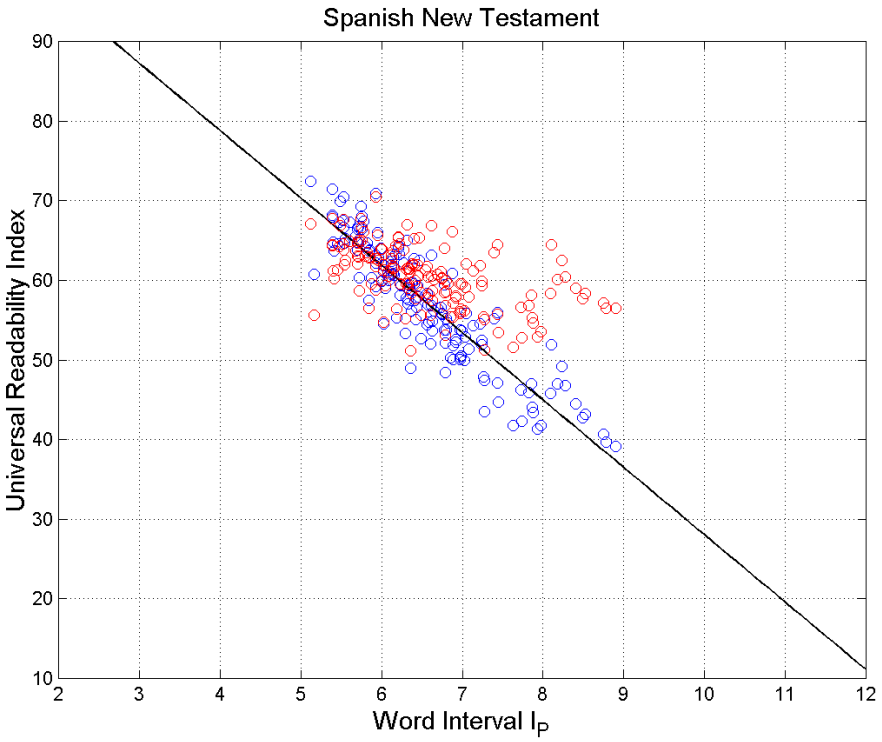


(**a**)

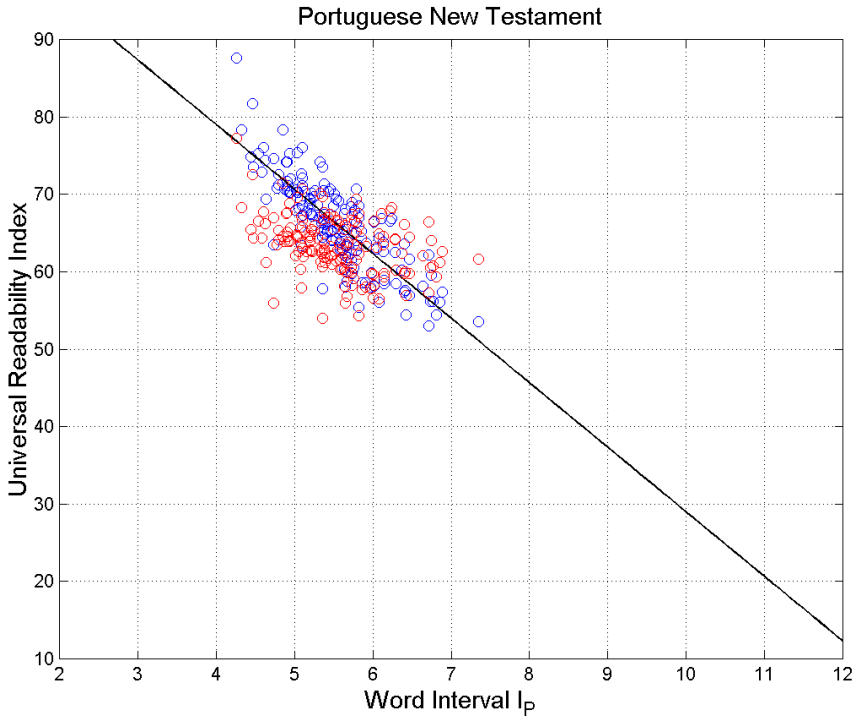


(**b**)

Figure A1. Scatter plot of the readability index G_U versus the word interval I_P (blue circles). The red circles refer to the scatterplots of G versus I_P. (**a**): Greek; (**b**): Latin. Miller's bounds are given by $I_P = 7 \mp 2$. The values of a and b of Equation (12) and the correlation coefficient between G_U and I_P are reported in Table A1.

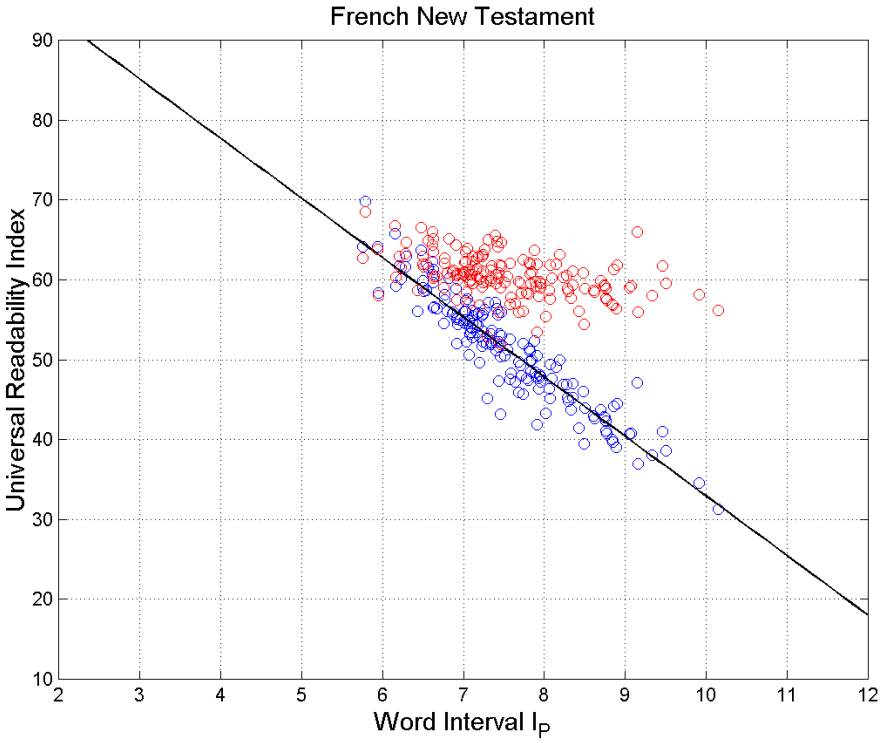


(**a**)

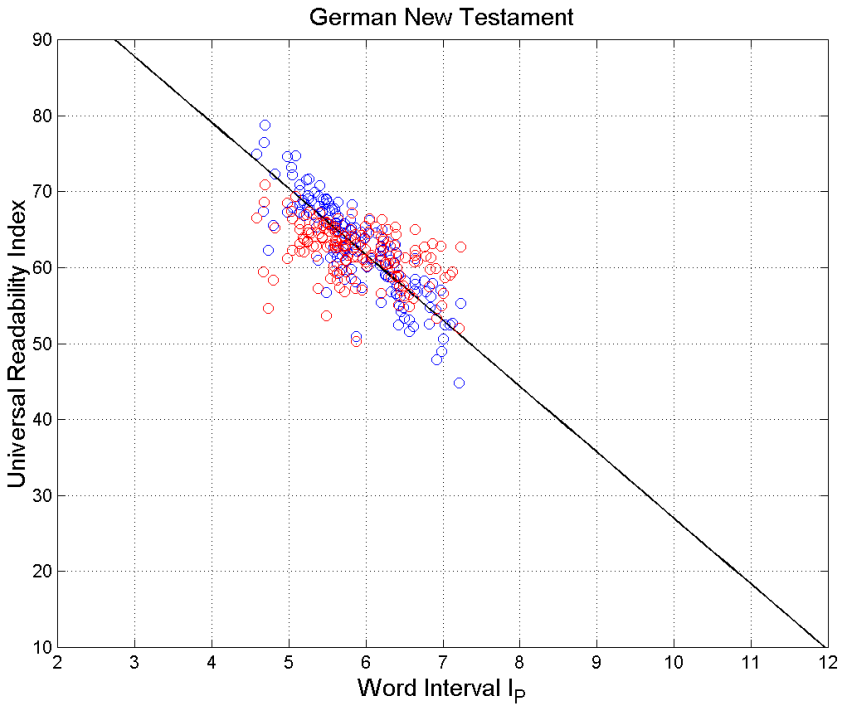


(**b**)

Figure A2. Scatter plot of the readability index G_U versus the word interval I_P (blue circles). The red circles refer to the scatterplots of G versus I_P. (**a**): Spanish; (**b**): Portuguese. Miller's bounds are given by $I_P = 7 \mp 2$. The values of a and b of Equation (12) and the correlation coefficient between G_U and I_P are reported in Table A1.



(**a**)



(**b**)

Figure A3. Scatter plot of the readability index G_U versus the word interval I_P (blue circles). The red circles refer to the scatterplots of G versus I_P. (**a**): French; (**b**): German. Miller's bounds are given by $I_P = 7 \mp 2$. The values of a and b of Equation (13) and the correlation coefficient between G_U and I_P are reported in Table A1.

Appendix B

Table A1. Values of a and b of Equation (12), and correlation coefficient between G_U and I_P for the indicated languages [11].

Language	a	b	Correlation Coefficient
Greek	8.62	113.66	−0.9477
Latin	10.59	120.82	−0.8666
Esperanto	9.87	114.20	−0.8803
French	7.46	107.51	−0.9311
Italian	7.80	108.54	−0.9065
Portuguese	8.34	112.33	−0.8261
Romanian	8.08	111.11	−0.8163
Spanish	8.46	112.60	−0.9061
Danish	9.46	120.71	−0.9182
English	7.88	110.23	−0.9129
Finnish	10.06	118.22	−0.8057
German	8.68	113.79	−0.8563
Icelandic	8.68	114.98	−0.8848
Norwegian	7.32	110.28	−0.9426
Swedish	7.32	109.98	−0.9546
Bulgarian	9.00	117.63	−0.8697
Czech	10.41	125.50	−0.8269
Croatian	9.86	122.33	−0.8868
Polish	9.98	123.60	−0.7160
Russian	10.70	118.04	−0.7326
Serbian	8.71	117.24	−0.8312
Slovak	10.03	124.83	−0.8417
Ukrainian	8.34	113.42	−0.7092
Estonian	9.97	120.11	−0.8643
Hungarian	10.83	118.91	−0.8034
Albanian	8.01	107.04	−0.8776
Armenian	12.11	133.87	−0.7805
Welsh	7.74	103.12	−0.7828
Basque	9.99	117.48	−0.8361
Hebrew	10.27	129.58	−0.8163
Cebuano	6.97	107.50	−0.9683
Tagalog	7.78	112.54	−0.9188
Chichewa	8.40	118.76	−0.9325
Luganda	8.69	118.42	−0.8713
Somali	8.65	113.41	−0.9492
Haitian	8.25	115.41	−0.9132
Nahuatl	7.55	113.02	−0.9420
Overall	8.94 ± 1.22	116.00 ± 6.49	0.8681 ± 0.0661

References

1. Flesch, R. A New Readability Yardstick. *J. Appl. Psychol.* **1948**, *32*, 222–233. [CrossRef] [PubMed]
2. Flesch, R. *The Art of Readable Writing*; revised and enlarged edition; Harper & Row: New York, NY, USA, 1974.
3. Kincaid, J.P.; Fishburne, R.P.; Rogers, R.L.; Chissom, B.S. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) For Navy Enlisted Personnel*; Research Branch Report 8-75; Chief of Naval Technical Training, Naval Air Station: Memphis, TN, USA, 1975.
4. DuBay, W.H. *The Principles of Readability*; Impact Information: Costa Mesa, CA, USA, 2004.
5. Bailin, A.; Graftstein, A. The linguistic assumptions underlying readability formulae: A critique. *Lang. Commun.* **2001**, *21*, 285–301. [CrossRef]
6. DuBay, W.H. (Ed.) *The Classic Readability Studies*; Impact Information: Costa Mesa, CA, USA, 2006.
7. Zamanian, M.; Heydari, P. Readability of Texts: State of the Art. *Theory Pract. Lang. Stud.* **2012**, *2*, 43–53. [CrossRef]
8. Benjamin, R.G. Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty. *Educ. Psychol. Rev.* **2011**, *24*, 63–88. [CrossRef]
9. Collins-Thompson, K. Computational Assessment of Text Readability: A Survey of Past, in Present and Future Research, Recent Advances in Automatic Readability Assessment and Text Simplification. *ITL Int. J. Appl. Linguist.* **2014**, *165*, 97–135. [CrossRef]

10. Kandel, L.; Moles, A. Application de l'indice de Flesch à la langue française. *Cah. Etudes Radio-Télévis.* **1958**, *19*, 253–274.
11. Matricciani, E. A Statistical Theory of Language Translation Based on Communication Theory. *Open J. Stat.* **2020**, *10*, 936–997. [CrossRef]
12. Lucisano, P.; Piemontese, M.E. GULPEASE: Una formula per la predizione della difficoltà dei testi in lingua italiana. *Sc. Città* **1988**, *3*, 110–124.
13. Matricciani, E. Deep Language Statistics of Italian throughout Seven Centuries of Literature and Empirical Connections with Miller's 7 ∓ 2 Law and Short-Term Memory. *Open J. Stat.* **2019**, *09*, 373–406. [CrossRef]
14. Miller, G.A. The Magical Number Seven, Plus or Minus Two. Some Limits on Our Capacity for Processing Information. *Psychol. Rev.* **1955**, *62*, 343–352.
15. Matricciani, E. Linguistic Mathematical Relationships Saved or Lost in Translating Texts: Extension of the Statistical Theory of Translation and Its Application to the New Testament. *Information* **2022**, *13*, 20. [CrossRef]
16. Matricciani, E. Multiple Communication Channels in Literary Texts. *Open J. Stat.* **2022**, *12*, 486–520. [CrossRef]
17. Matricciani, E. Capacity of Linguistic Communication Channels in Literary Texts: Application to Charles Dickens' Novels. *Information* **2023**, *14*, 68. [CrossRef]
18. François, T. An analysis of a French as Foreign language corpus for readability assessment. In *Proceedings of the 3rd Workshop on NLP for CALL*; NEALT Proceedings Series 22; Linköping 2014 Electronic Conference Proceedings; Linköping University Electronic Press: Linköping, Sweden, 2014; Volume 107, pp. 13–32.
19. Baddeley, A.D.; Thomson, N.; Buchanan, M. Word Length and the Structure of Short-Term Memory. *J. Verbal Learn. Verbal Behav.* **1975**, *14*, 575–589. [CrossRef]
20. Cowan, N. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behav. Brain Sci.* **2000**, *24*, 87–114. [CrossRef]
21. Pothos, E.M.; Joula, P. Linguistic structure and short-term memory. *Behav. Brain Sci.* **2000**, *24*, 138–139. [CrossRef]
22. Jones, G.; Macken, B. Questioning short-term memory and its measurements: Why digit span measures long-term associative learning. *Cognition* **2015**, *144*, 1–13. [CrossRef]
23. Saaty, T.L.; Ozdemir, M.S. Why the Magic Number Seven Plus or Minus Two. *Math. Comput. Model.* **2003**, *38*, 233–244. [CrossRef]
24. Mathy, F.; Feldman, J. What's magic about magic numbers? Chunking and data compression in short-term memory. *Cognition* **2012**, *122*, 346–362. [CrossRef]
25. Chen, Z.; Cowan, N. Chunk Limits and Length Limits in Immediate Recall: A Reconciliation. *J. Exp. Psychol. Mem. Cogn.* **2005**, *31*, 1235–1249. [CrossRef]
26. Chekaf, M.; Cowan, N.; Mathy, F. Chunk formation in immediate memory and how it relates to data compression. *Cognition* **2016**, *155*, 96–107. [CrossRef] [PubMed]
27. Barrouillest, P.; Camos, V. As Time Goes By: Temporal Constraints in Working Memory. *Curr. Dir. Psychol. Sci.* **2012**, *21*, 413–419. [CrossRef]
28. Conway, A.R.A.; Cowan, N.; Michael, F.; Bunting, M.F.; Therriaulta, D.J.; Minkoff, S.R.B. A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence* **2002**, *30*, 163–183. [CrossRef]
29. Manzoni, A. *The Betrothed*; Moore, M.F., Translator; The Modern Library: New York, NY, USA, 2022.
30. Mazza, A. *Studi Sulle Redazioni de I Promessi Sposi*; Edizioni Paoline: Milan, Ialy, 1968.
31. Giovanni Nencioni, N. *La Lingua di Manzoni. Avviamento Alle Prose Manzoniane*; Il Mulino: Bologna, Italy, 1993.
32. Guntert, G. *Manzoni Romanziere: Dalla Scrittura Ideologica Alla Rappresentazione Poetica*; Franco Cesati Editore: Firenze, Italy, 2000.
33. Frare, P. *Leggere I Promessi Sposi*; Il Mulino: Bologna, Italy, 2016.