

Deep Language Statistics of Italian throughout Seven Centuries of Literature and Empirical Connections with Miller's 7 ± 2 Law and Short-Term Memory

Emilio Matricciani

Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano, Milan, Italy

Email: Emilio.Matricciani@polimi.it

How to cite this paper: Matricciani, E. (2019) Deep Language Statistics of Italian throughout Seven Centuries of Literature and Empirical Connections with Miller's 7 ± 2 Law and Short-Term Memory. *Open Journal of Statistics*, 9, 373-406. <https://doi.org/10.4236/ojs.2019.93026>

Received: May 29, 2019

Accepted: June 27, 2019

Published: June 30, 2019

Copyright © 2019 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Statistics of languages are usually calculated by counting characters, words, sentences, word rankings. Some of these random variables are also the main “ingredients” of classical readability formulae. Revisiting the readability formula of Italian, known as GULPEASE, shows that of the two terms that determine the readability index G —the *semantic index* G_C , proportional to the number of characters per word, and the *syntactic index* G_F , proportional to the reciprocal of the number of words per sentence— G_F is dominant because G_C is, in practice, constant for any author throughout seven centuries of Italian Literature. Each author can modulate the length of sentences more freely than he can do with the length of words, and in different ways from author to author. For any author, any couple of text variables can be modelled by a linear relationship $y = mx$, but with different slope m from author to author, except for the relationship between characters and words, which is unique for all. The most important relationship found in the paper is that between the short-term memory capacity, described by Miller's “ 7 ± 2 law” (*i.e.*, the number of “chunks” that an average person can hold in the short-term memory ranges from 5 to 9), and the *word interval*, a new random variable defined as the average number of words between two successive punctuation marks. The word interval can be converted into a *time interval* through the average reading speed. The *word interval* spreads in the same range as Miller's law, and the *time interval* is spread in the same range of short-term memory response times. The connection between the word interval (and time interval) and short-term memory appears, at least empirically, justified and natural, however, to be further investigated. Technical and scientific writings (papers, essays, etc.) ask more to their

readers because words are on the average longer, the readability index G is lower, word and time intervals are longer. Future work done on ancient languages, such as the classical Greek and Latin Literatures (or modern languages Literatures), could bring us an insight into the short-term memory required to their well-educated ancient readers.

Keywords

GULPEASE, Italian, Literature, Miller's Law, Readability, Short-Term Memory, Word Interval

1. Introduction

Statistics of languages have been calculated for several western languages, mostly by counting characters, words, sentences, word rankings [1]. Some of these parameters are also the main “ingredients” of classical readability formulae. First developed in the United States [2], readability formulae are applicable to any language, once the mathematical expression for that particular language is developed and assessed experimentally [3]. Readability formulae measure textual characteristics that are quantifiable, therefore mainly words and sentences lengths, by defining an ad hoc mathematical index. Therefore, according to the classical readability formulae, and solely on this ground, different texts can be compared automatically to assess the difficulty a reader should tolerate before giving up, if he is allowed to do so when he reads for pleasure not for duty, as is the case with technical and scientific texts [4] [5] [6], which must be read because of technical or research activities, or for studying. In other words, readability indices allow matching texts to expected readers to the best possible, by avoiding over difficulty and inaccessible texts, or oversimplification, the latter felt as making fun of the reader.

Even after many years of studies and proposals of many readability formulae, especially for English [3] [7] [8] [9], nobody, as far as I know, has shown a possible direct relationship of a readability formula and its constituents (characters, sentences etc.), with reader's short-term memory behavior [10] [11] [12] [13]. In this paper, I show that a statistical relationship can be found for Italian, by examining a large number of literary texts written since the XIV century, thus revitalizing the classical readability formula approach.

The classical readability formulae, in fact, have been criticized because they focus on a limited set of superficial text features, rough approximations of the linguistic factors at play in readability assessment [14]. However, a readability formula does measure important constituents of texts and can contribute to understanding the process of communication, especially if its ingredients relate to the storage of information in the short-term memory. Moreover, because an “absolute” (*i.e.* a formula that provides numerical indices counted from a universal origin, such as zero) readability formula might not exist at all, the current

formulae can be used to compare different texts, together with other parameters which I define in this paper. In other words, differences between the numerical values given by a readability formula—*i.e.* relative indices—may be more significant than absolute ones for the purpose of comparing texts, especially literary texts, as done in this paper. In spite of this, in the following I will present absolute values because readers are used to considering them when they apply readability formulae, but I also discuss differences, which the reader can appreciate from the numerical values reported in the tables or in the scatter plots below.

In spite of the long-lasting controversy on the development and use of classical readability formulae, researchers still continue to develop methods to overcome weaknesses by advancing natural language processing and other computerized language methods [15] [16], to capture more complex linguistic features [17]. Benjamin [15], however, predicts that also in this research field will happen what does happen in any research field when no general consensus is shared on a specific topic, that is, that also these new developments will be judged controversial. In any case, the classical readability formulae have served their purpose in leveling typical books for schoolchildren and general audience, such as in Italy in the 1980s [18].

Now, new methods, developed after cognitive processing theories, should allow analyzing more complex texts for specific targets such as adolescents, university students, and adults. Moreover, with machine-learning developments, non-traditional texts, like those found in many web sites, can be categorized for greater accessibility. Some of these advances concern even observing eye tracking while reading [16] [19]. For Italian, the work by Dell’Orletta and colleagues [17] aims at automatically assessing the readability of newspaper texts with the specific task of text simplification, not for specifically analyzing and studying literary texts and their statistics, as I do in this paper.

A readability formula is, however, very attractive because it allows giving a quantitative and automatic judgement on the difficulty or easiness of reading a text. Every readability formula, however, gives a *partial* measurement of reading difficulty because its result is mainly linked to words and sentences length. It gives no clues as to the correct use of words, to the variety and richness of the literary expression, to its beauty or efficacy, does not measure the quality and clearness of ideas or give information on the correct use of grammar, does not help in better structuring the outline of a text, for example a scientific paper. The *comprehension* of a text (not to be confused with its readability, defined by the mathematical formulae) is the result of many other factors, the most important being reader’s culture and reading habits. In spite of these limits, readability formulae are very useful, if we apply them for specific purposes, and assess their possible connections with the short-term memory of readers.

Compared to the more sophisticated methods mentioned above the classical readability formulae have several advantages:

- 1) They give an index that any writer (or reader) can calculate directly, easily, by

means of the same tool used for writing (e.g. WinWord), therefore sufficiently matching the text to the expected audience.

- 2) Their “ingredients” are understandable by anyone, because they are interwound with a long-lasting writing and reading experience based on characters, words and sentences.
- 3) Characters, words, sentences and punctuation marks appear to be related to the capacity and time response of short-term memory, as shown in this paper.
- 4) They give an index based on the same variables, regardless of the text considered, thus they give an objective measurement for comparing different texts or authors, without resorting to readers’ physical actions or psychological behavior, which largely vary from one reader to another, and within a reader in different occasions, and may require ad hoc assessment methods.
- 5) A final objective readability formula or more recent software-developed methods valid universally are very unlikely to be found or accepted by everyone. Instead of absolute readability, readability differences can be more useful and meaningful. The classical readability formulae provide these differences easily and directly.

In this paper, for Italian, I show that a relationship between some texts statistics and reader’s short-term memory capacity and response time seems to exist. I have found an empirical relationship between the readability formula mostly used for Italian and short-term memory capacity, by considering a very large sample of literary works of the Italian Literature spanning seven centuries, most of them still read and studied in Italian high schools or researched in universities. The contemporaneous reader of any of these works is supposed to be, of course, educated and able to read long texts with good attention. In other words, this audience is quite different of that considered in the studies and experiments reported above on new techniques (based on complex software) for assessing readability of specific types of texts [17]. In other words, the subject of my study are the ingredients of a classical readability formula, not the formula itself (even though I have found some interesting features and limits of it), and its empirical relationship with short-term memory. From my results, it might be possible to establish interesting links to other cognitive issues, as discussed by [20], a task beyond the scope of this paper and author’s expertise.

The most important relationship I have found is that between the short-term memory capacity, described by Miller’s “ 7 ± 2 law” [21], and what I call the *word interval*, a new random variable defined as the average number of words between two successive punctuation marks. The word interval can be converted into a *time interval* through the average reading speed. The *word interval* is numerically spread in a range very alike to that found in Miller’s law, and more recently by Jones and Macken [12], and the *time interval* is spread in a range very alike to that found in the studies on short-term memory response time [10] [22] [23]. The connection between the word interval (and time interval) and

short-term memory appears, at least empirically, justified and natural, but at this stage is a well-educated guess.

Finally, notice that in the case of ancient languages, no longer spoken by a people but rich in literary texts, such as Greek or Latin, that have founded the Western civilization, it is obvious that nobody can make reliable experiments, as those reported in the references recalled above. These ancient languages, however, have left us a huge library of literary and (few) scientific texts. Besides the traditional count of characters, words and sentences, the study of word and time intervals statistics should bring us an insight into the short-term memory features of these ancient readers, and this can be done very easily, as I have done for Italian. An analysis of the New Testament Greek originals [24], similar to that reported in this paper, shows results very similar to those reported in this paper, therefore evidencing some universal and long-lasting characteristics of western languages and their readers.

In conclusion, the aim of this paper is to research, with regard to the high Italian language, the following topics:

- 1) The impact of character and sentence indices on the readability index (all defined in Section 2).
- 2) The relationship of these indices with the newly defined “word interval” and “time interval”.
- 3) The “distance”, absolute and relative, of literary texts by defining meaningful vectors based on characters, words, sentences, punctuation marks.
- 4) The relationship between the word interval and Miller’s law, and between the time interval and short-term memory response time.

After this Introduction, Section 2 revisits the classical readability formula of Italian; Section 3 shows interesting relationships between its constituents; Section 4 reports the statistical results for a large number of texts of the Italian Literature since the XIV century; Section 5 discusses the “distance” of literary texts; Section 6 introduces word and time intervals and their empirical relationships with short-term memory features; Section 7 discusses some different results concerning scientific and technical texts, and finally Section 8 draws some conclusions and suggests future work.

2. Revisiting the GULPEASE Readability Formula of Italian

For Italian, the most used formula (calculated by WinWord, for example), known with the acronym GULPEASE [25], is given by:

$$G = 89 - 10 \times \frac{c}{p} + 300 \times \frac{f}{p} \quad (1a)$$

The numerical values of Equation (1a) can be interpreted as readability index for Italian as a function of the number of years of school attended in Italy’s school system [25], as I have graphically summarized in **Figure 1**. The larger G , the more readable the text is. In Equation (1a) p is the total number of words in the text considered, c is the number of letters contained in the p words, f is the

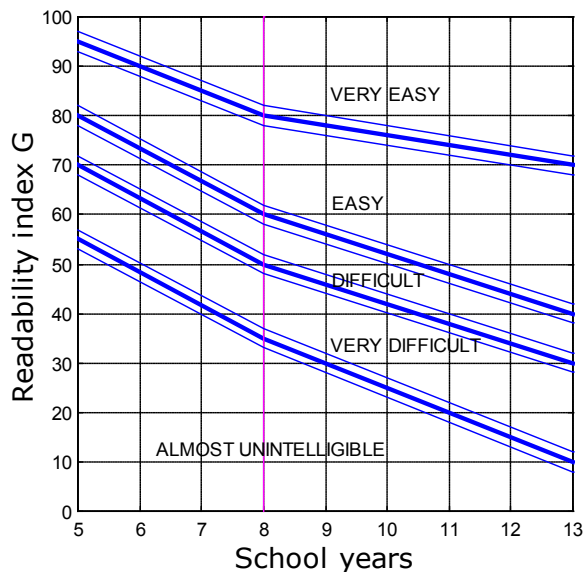


Figure 1. Readability index G of Italian, as a function of the number of school years attended (in Italy high school lasts 5 years, children attend it up to 19 years old). The blue thinner lines indicate the error bounds found in using Equation (6). Elementary school lasts 5 years, Scuola Media Inferiore lasts 3 years, Scuola Media Superiore (High School) lasts 5 years (the vertical magenta line shows the beginning of High School).

number of sentences contained in the p words (a list of mathematical symbols is reported in the Appendix). Let us define the terms:

$$G_C = 10 \times \frac{c}{p} \tag{2a}$$

$$G_F = 300 \times \frac{f}{p} \tag{2b}$$

Therefore, Equation (1a) can be written as:

$$G = 89 - G_C + G_F \tag{1b}$$

We analyze first Equations (1a) (1b), by means of standard statistics of its addends, because, as other readability formulae, they contain important characteristics of literary texts which for the Italian Literature, that extends for the longest period of time compared to other modern western languages, have been stable over centuries (namely G_C).

Equation (1) says that a text, for the same amount of words, is more difficult to read if f/p is small, hence if sentences are long, and if the number of characters per word $C_p = c/p$ is large, hence if words are long. Long sentences mean that the reciprocal value $P_F = \frac{p}{f} = \frac{300}{G_F}$ is large, therefore there are many

words in a sentence, G_F decreases and thus G decreases. The sentences contain many subordinate clauses, the reading difficulty is due to syntax and therefore we term loosely G_F the *syntactic index*. Long words mean that G_C increases, it is subtracted from the constant 89 and thus G decreases. Long words often refer

to abstract concepts (maybe of Latin or Greek origin), difficulty is due to characters, and therefore we term loosely G_C the *semantic index*. In other words, a text is easier to read if it contains short words and short sentences, a known result applicable to readability formulae of any language.

Now, the study of Equation (1), and in particular how the two terms G_C , G_F affect the value of G , brings very interesting results, as we show next. In this paper I apply the above equations to classical literary works of a large number of Italian writers¹, from Giovanni Boccaccio (XIV century) to Italo Calvino (XX century), see **Table 1**, by examining some complete works, as they are available today in their best edition².

Table 1. Characters, words and sentences in the literary works considered in this study, and average values of the corresponding G , G_C e G_F , the standard deviation of averages (σ_μ in parentheses) and the standard deviation σ_r estimated for text blocks of 1000 words³. The characters are those contained in the words. All parameters have been computed by weighting the text blocks according to the number of the words contained in them. For instance, in *Decameron*, the average value of G can be estimated in 51.18 ± 0.17 and its standard deviation for text blocks of 1000 words is 2.85.

Author, literary work, century	Characters	Words	Sentences	G	G_C	G_F
Anonymous (<i>I Fioretti di San Francesco</i> , XIV)	180,056	38,681	1064	50.70	46.55	8.25
				(0.30)	(0.163)	(0.24)
Bembo Pietro (<i>Prose</i> , XV-XVI)	295,614	67,572	1925	1.84	1.01	1.48
				53.80	43.75	8.55
Bembo Pietro (<i>Prose</i> , XV-XVI)	295,614	67,572	1925	(0.66)	(0.30)	(0.44)
				5.42	2.44	3.65
Boccaccio Giovanni (<i>Decameron</i> , XIV)	1,190,417	266,033	6147	51.18	44.75	6.94
				(0.17)	(0.11)	(0.10)
Boccaccio Giovanni (<i>Decameron</i> , XIV)	1,190,417	266,033	6147	2.85	1.84	1.63
				55.27	51.04	17.30
Buzzati Dino (<i>Il deserto dei tartari</i> , XX)	292,974	57,402	3311	(0.54)	(0.27)	(0.51)
				4.08	2.03	3.86
Buzzati Dino (<i>La boutique del mistero</i> , XX)	302,894	62,771	4219	60.91	48.25	20.16
				(0.81)	(0.19)	(0.69)
Buzzati Dino (<i>La boutique del mistero</i> , XX)	302,894	62,771	4219	6.40	1.50	5.44
				58.93	46.32	16.25
Calvino (<i>Il barone rampante</i> , XX)	330,420	71,340	3864	(0.84)	(0.23)	(0.77)
				7.10	1.91	6.53
Calvino Italo (<i>Marcovaldo</i> , XX)	161,952	34,206	2000	59.19	47.35	17.54
				(0.85)	(0.24)	(0.74)
Calvino Italo (<i>Marcovaldo</i> , XX)	161,952	34,206	2000	4.99	1.42	4.30

¹Information about authors and their literary texts can be found in any history of Italian literature, or in dictionaries of Italian literature.

²The great majority of these texts are available in digital format at <https://www.liberliber.it>.

³The standard deviation found in n text blocks $\sigma = \sqrt{v - \mu^2}$ is scaled to a reference text of $p_r = 1000$ words by first calculating the number of text blocks with this length, namely $n_r = p_r/p$, and then scaling σ as $\sigma_r = \sigma \times \sqrt{\frac{n_r}{n}} = \sigma_\mu \sqrt{n_r}$.

Continued

Cassola Carlo (<i>La ragazza di Bube</i> , XX)	307,819	68,698	5873	69.84 (1.11) 9.22	44.81 (0.22) 1.84	25.65 (0.96) 7.95
Collodi Carlo (<i>Pinocchio</i> , XIX)	186,849	40,642	2512	61.57 (0.79) 5.04	45.97 (0.22) 1.41	18.54 (0.70) 4.47
Da Ponte Lorenzo (<i>Vita</i> , XVIII-XIX)	646,024	137,054	5459	53.81 (0.50) 5.87	47.14 (0.14) 1.65	11.95 (0.43) 5.04
Deledda Grazia (<i>Canne al vento</i> , XX, Nobel Prize 1926)	276,552	61,375	4184	64.39 (0.92) 7.21	45.06 (0.18) 1.39	20.45 (0.79) 6.22
D'Azeglio Massimo (<i>Ettore Fieramosca</i> , XIX)	424,259	91,464	3182	53.05 (0.54) 5.15	46.39 (0.17) 1.62	10.44 (0.45) 4.33
De Amicis Edmondo (<i>Cuore</i> , XIX)	376,792	82,770	4775	60.78 (0.55) 5.01	45.52 (0.15) 1.39	17.31 (0.54) 4.94
De Marchi Emilio (<i>Demetrio Panelli</i> , XIX)	471,451	100,328	5363	58.05 0.44 4.37	46.99 0.14 1.43	16.04 0.35 3.48
D'Annunzio Gabriele (<i>Le novelle delle Pescara</i> , XX)	244,055	49,688	3027	58.16 (0.81) 5.74	49.12 (0.16) 1.12	18.28 (0.77) 5.44
Eco Umberto (<i>Il nome della rosa</i> , XX)	821,707	170,676	8490	55.78 (0.52) 6.82	48.14 (0.11) 1.47	14.92 (0.49) 6.35
Fogazzaro (<i>Il santo</i> , XIX-XX)	467,990	97,616	6637	61.46 (0.72) 7.16	47.94 (0.10) 1.02	20.40 (0.65) 6.40
Fogazzaro (<i>Piccolo mondo antico</i> , XIX-XX)	528,659	112,706	7069	61.46 (0.57) 6.06	47.94 (0.16) 1.65	20.40 (0.48) 5.12
Gadda (<i>Quer pasticciaccio brutto...</i> XX)	474,359	99,631	5596	58.24 (1.08) 10.77	47.61 (0.24) 2.35	16.85 (1.00) 9.99
Grossi Tommaso (<i>Marco Visconti</i> , XIX)	637,311	138,900	5301	54.57 0.63 7.39	45.88 0.12 1.42	11.45 0.55 6.53
Leopardi Giacomo (<i>Operette morali</i> , XIX)	322,991	68,699	2694	53.76 (1.17) 8.46	47.01 (0.32) 2.62	11.77 (0.96) 6.46
Levi (<i>Cristo si è fermato a Eboli</i>)	383,893	81,092	3611	55.02 (0.39) 3.54	47.34 (0.13) 1.13	13.36 (0.36) 3.27
Machiavelli Niccolò (<i>Il principe</i> , XV-XVI)	130,274	27,680	702	49.54 (0.33) 1.75	47.06 (0.21) 1.10	7.61 (0.21) 1.09

Continued

Manzoni Alessandro (<i>I promessi sposi</i> , XIX)	1,036,728	225,392	9766	56.00 (0.69) 10.29	46.00 (0.21) 3.18	13.00 (0.52) 7.76
Manzoni Alessandro (<i>Fermo e Lucia</i> , XIX)	1,044,997	219,993	7496	51.72 (0.53) 2.84	47.50 (0.19) 2.84	10.22 (0.37) 5.54
Moravia Alberto (<i>Gli indifferenti</i> , XX)	471,574	98,084	2830	49.58 (0.40) 1.48	48.08 (0.15) 1.48	8.66 (0.40) 4.00
Moravia Alberto (<i>La ciociara</i> , XX)	577,176	126,550	4271	53.52 (0.38) 4.27	45.61 (0.24) 2.70	10.12 (0.33) 3.68
Pavese Cesare (<i>La bella estate</i> , XX)	116,360	25,650	2121	68.44 (0.90) 4.54	45.36 (0.17) 0.88	24.81 (0.85) 4.29
Pavese Cesare (<i>La luna e i falò</i> , XX)	194,032	43,442	2544	61.90 (0.65) 4.27	44.66 (0.14) 0.93	17.57 (0.66) 4.36
Pellico Silvio (<i>Le mie prigioni</i> , XIX)	252,915	52,644	3148	58.90 (0.37) 2.65	48.04 (0.12) 0.87	17.94 (0.32) 2.30
Pirandello Luigi (<i>Il fu Mattia Pascal</i> , Nobel Prize 1934)	345,301	74,544	5284	63.94 (1.01) 8.71	46.32 (0.20) 1.69	21.26 (0.93) 8.02
Sacchetti Franco (<i>Trecentonovelle</i> , XIV)	767,538	175,452	8060	59.04 (0.40) 5.35	43.75 (0.12) 1.62	13.78 (0.35) 4.63
Salernitano Masuccio (<i>Il Novellino</i> , XV)	152,345	34,623	1965	62.03 (0.89) 5.25	44.00 (0.20) 1.18	17.03 (0.91) 5.36
Salgari Emilio (<i>Il corsaro nero</i> , XIX-XX)	493,213	98,945	6686	59.42 (0.51) 5.09	49.85 (0.12) 1.21	20.27 (0.46) 4.58
Salgari Emilio (<i>I minatori dell'Alaska</i> , XIX-XX)	453,614	90,486	6094	59.07 (0.55) 5.22	50.13 (0.12) 1.14	20.20 (0.54) 5.11
Svevo Italo (<i>Senilità</i> , XX)	325,221	66,912	4236	59.39 (0.65) 5.33	48.60 (0.10) 0.81	19.00 (0.58) 4.78
Tomasi di Lampedusa (<i>Il gattopardo</i> , XX)	371,853	74,462	2893	50.72 (0.81) 6.96	49.94 (0.20) 1.73	11.66 (0.71) 6.09
Verga (<i>I Malavoglia</i> , XIX-XX)	393,902	88,277	4401	59.34 (0.56) 5.31	44.62 (0.51) 1.45	14.96 (0.15) 4.82
Global values	16,502,125	3,533,155	169,636	56.71 (0.16)	46.70 (0.06)	14.40 (0.15)

3. Relationships among G_C , G_F and G

The semantic index G_C , given by the number of characters per word multiplied by 10 (Equation (2a)), and the syntactic index G_P , given by the reciprocal of the number of words per sentence P_P , multiplied by 300 (Equation (2b)), affect very differently the final value of G (Equation (1b)). **Table 1** lists the average values of G , G_C , G_F and their standard deviations for the literary works considered. In this analysis, as in the successive ones, I have considered text blocks, singled out by an explicit subdivision of the author or editor (e.g., chapters, subdivision of chapters, etc.), without titles. This arbitrary selection does not affect average values and the standard deviations of these averages. All parameters have been calculated by weighting any text block with its number of words, so that longer blocks weigh statistically more than shorter ones as detailed in the following.

Let n be the number of text blocks contained in a literary work and $p_T = \sum_{i=1}^n p_i$ the total number of words in it. The average value μ and the standard deviation of the average value σ_μ of each parameter are calculated by weighing each text block with its the number of words. For example, for G_F the average value is given by $\mu = 300 \times \sum_{i=1}^{i=n} \frac{f_i}{p_i} \times \frac{P_i}{p_T}$ with p_i , f_i the number of

words and sentences contained in the text block i -th. For the standard deviation of the average value, we calculate first the average square value

$$v = 300^2 \sum_{i=1}^{i=n} \left(\frac{f_i}{p_i} \right)^2 \times \frac{P_i}{p_T}$$

$$\sigma = \sqrt{v - \mu^2}, \text{ and finally we calculate } \sigma_\mu = \frac{\sigma}{\sqrt{n}}.$$

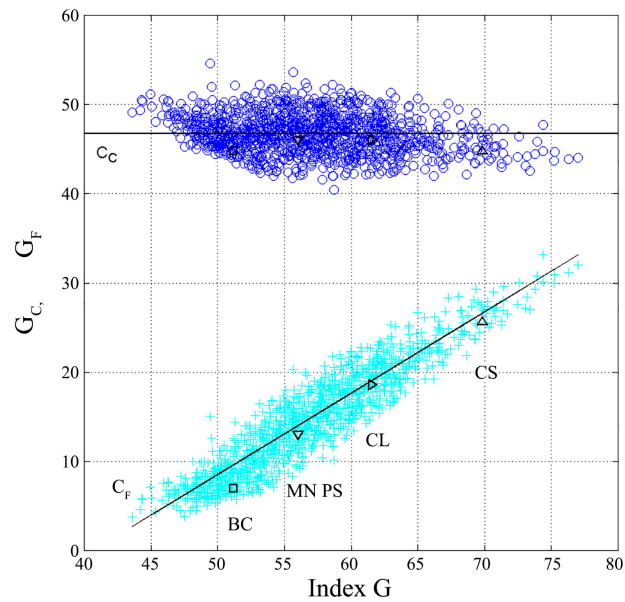
In this way different literary works can be reliably compared with regard to any parameter, regardless of the choice of the length of text blocks.

From the results reported in **Table 1**, it is evident that G_C changes much less than G_P , a feature highlighted in the scatter plot of **Figure 2(a)**, which shows G_C and G_F versus G , for each text block (1260 text blocks in total, with different number of words) found in the listed literary works.

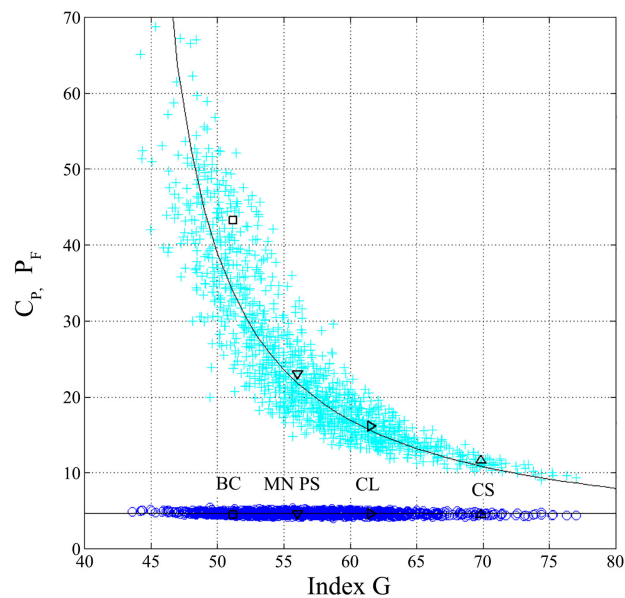
The theoretical range of G can be calculated by considering the theoretical range of G_P . The maximum value of G_P is found when P_P is minimum, the latter given by 1 when $f = p = 1$, therefore when all sentences are made of 1 single word, hence $G_{F,\max} = 300$, a case obviously not realistic. A more realistic maximum value can be estimated by considering 4 or 5 words per sentence, so that $G_{F,\max}$ reduces to 75 or 60. The minimum value is obviously $G_{F,\min} = 0$, *i.e.*, the text is made of 1 sentence with an infinite (very large) number of words. In conclusion, by considering the average value of $G_C = 46.7$ (**Table 1**), the GULPEASE index can theoretically range from $G_{\max} = 89 - 46.7 + 60 = 102.3$ to $G_{\min} = 89 - 46.7 + 0 = 42.3$ (close to the smallest values in **Table 1**).

The constancy of G_C versus G indicates that, in Italian, the number of characters per word C_P has been very stable over many centuries, while the linear proportionality between G_F and G , is directly linked to author's style, or to the style

applied to different works by the same author. With “style” I refer to the writer’s choice of sentence length and punctuation marks distribution within the sentence, namely to the variables of interest in my mathematical approach. These features are confirmed in **Figure 2(b)**, which shows the scatter plots of the average number of characters per word C_p vs. G , and P_f vs. G . In other words, the



(a)



(b)

Figure 2. (a) G_C (blue circles) and G_F (cyan crosses) versus G , for all 1260 text blocks. BC = Boccaccio, MN PS = Manzoni (*I promessi sposi*), CL = Collodi, CS = Cassola. The horizontal line is the average value of G_C , Equation (3), the diagonal line is the average value of G_F , Equation (4); (b) Scatter plots of $C_p = c/p$ vs. G (blue dots) and $P_f = p/f$ vs. G (cyan crosses). BC = Boccaccio, MN PS = Manzoni (*I promessi sposi*), CL = Collodi, CS = Cassola. The black continuous lines are given by Equations (3) and (5).

readability of a text using (1) is practically due only to the syntactic index G_p therefore to the number of words per sentence. The two lines drawn in **Figure 2(a)** are given by the average value of G_C (**Table 2**):

Table 2. Average values of a number of characters per word, words per sentence, punctuation marks per sentence and punctuation interval. Standard deviations calculated as in **Table 1**.

Author	Characters per word C_p	Words per sentence P_F	Punctuation marks per sentence M_F	Words per punctuation mark (word interval I_p)
Anonymous	4.65 (0.02) 0.10	37.70 (1.19) 7.39	4.56 (0.12) 0.74	8.24 (0.11) 0.67
Bembo	4.37 (0.03) 0.24	37.91 (2.16) 17.72	5.92 (0.25) 1.98	6.42 (0.34) 1.02
Boccaccio	4.48 (0.01) 0.18	44.27 (0.59) 9.69	5.69 (0.07) 1.13	7.79 (0.06) 0.92
Buzzati (D)	5.10 (0.03) 0.20	17.75 (0.50) 3.81	2.67 (0.06) 0.43	6.63 (0.11) 0.86
Buzzati (B)	4.82 (0.02) 0.15	15.45 (0.62) 4.91	2.41 (0.06) 0.45	6.37 (0.15) 1.16
Calvino (B)	4.63 (0.02) 0.19	19.87 (1.02) 8.61	2.91 (0.10) 0.85	6.73 (0.14) 1.18
Calvino (M)	4.74 (0.02) 0.14	17.60 (0.64) 3.74	2.67 (0.08) 0.49	6.59 (0.14) 0.79
Cassola	4.48 (0.02) 0.18	11.93 (0.46) 3.80	2.11 (0.05) 0.39	5.64 (0.10) 0.86
Collodi	4.60 (0.02) 0.14	16.92 (0.60) 3.83	2.72 (0.08) 0.48	6.19 (0.08) 0.51
Da Ponte	4.71 (0.01) 0.17	26.15 (0.99) 11.55	3.78 (0.13) 1.57	6.91 (0.08) 0.93
Deledda	4.51 (0.02) 0.139	15.08 (0.64) 5.03	2.48 (0.06) 0.47	6.06 (0.16) 1.24
D'Azeglio	4.64 (0.02) 0.162	29.77 (1.21) 11.52	4.03 (0.13) 1.25	7.36 (0.11) 1.01
De Amicis	4.55 (0.02) 0.14	19.43 (0.74) 6.76	3.41 (0.11) 0.97	5.61 (0.06) 0.58
De Marchi	4.70 (0.01) 0.43	18.95 (0.40) 4.01	2.68 (0.05) 0.46	7.06 (0.06) 0.64
D'Annunzio	4.91 (0.16) 0.112	17.99 (0.79) 5.57	2.79 (0.08) 0.57	6.38 (0.16) 1.15
Eco	4.81 (0.01) 0.15	21.08 (0.65) 8.50	2.81 (0.07) 0.85	7.46 (0.10) 1.32
Fogazzaro (S)	4.79 (0.01) 0.10	14.84 (0.44) 4.32	2.34 (0.05) 0.53	6.33 (0.10) 0.94

Continued

Fogazzaro (P)	4.79 (0.02) 0.17	16.08 (0.42) 4.43	2.64 (0.06) 0.65	6.10 (0.10) 1.01
Gadda	4.76 (0.02) 0.24	18.43 (1.06) 10.60	3.68 (0.17) 1.70	4.98 (0.09) 0.86
Grossi	4.59 (0.01) 0.14	28.07 (1.26) 14.85	4.23 (0.13) 1.55	6.56 (0.13) 1.55
Leopardi	4.70 (0.03) 2.62	31.78 (2.05) 16.55	4.54 (0.25) 2.02	6.90 (0.23) 1.20
Levi	4.73 (0.01) 0.11	22.94 (0.61) 5.53	4.02 (0.11) 0.96	5.70 (0.03) 0.31
Machiavelli	4.71 (0.02) 0.11	40.17 (1.02) 5.39	6.23 (0.15) 0.81	6.45 (0.08) 0.41
Manzoni (PS)	4.60 (0.02) 0.32	24.83 (1.18) 17.79	4.63 (0.18) 2.70	5.30 (0.06) 0.95
Manzoni (FL)	4.75 (0.02) 0.28	30.98 (1.28) 19.03	4.30 (0.14) 2.10	7.17 (0.14) 2.05
Moravia (I)	4.81 (0.02) 0.15	36.00 (1.88) 18.61	5.34 (0.27) 2.64	6.74 (0.09) 0.93
Moravia (C)	4.56 (0.02) 0.27	29.93 (0.87) 9.73	4.12 (0.11) 1.20	7.28 (0.19) 2.13
Pavese (B)	4.54 (0.02) 0.09	12.37 (0.48) 2.43	2.06 (0.05) 0.25	5.97 (0.11) 0.56
Pavese (F)	4.47 (0.01) 0.93	17.83 (0.67) 4.42	2.60 (0.08) 0.53	6.83 (0.13) 0.83
Pellico	4.80 (0.01) 0.087	17.27 (0.33) 2.40	2.69 (0.06) 0.41	6.50 (0.08) 0.61
Pirandello	4.63 (0.02) 0.169	14.57 (0.62) 5.35	2.93 (0.08) 0.69	4.94 (0.10) 0.86
Sacchetti	4.37 (0.01) 0.16	22.43 (0.58) 7.66	3.83 (0.06) 0.82	5.82 (0.08) 1.01
Salernitano	4.40 (0.02) 0.12	19.20 (1.18) 6.96	3.68 (0.17) 1.03	5.14 (0.08) 0.46
Salgari (C)	4.99 (0.01) 0.121	15.09 (0.37) 3.63	2.36 (0.04) 0.38	6.36 (0.07) 0.70
Salgari (M)	5.01 (0.01) 0.11	15.24 (0.40) 3.85	2.44 (0.05) 0.45	6.246 (0.08) 0.74
Svevo	4.86 (0.001) 0.08	16.04 (0.58) 4.76	2.07 (0.07) 0.59	7.75 (0.13) 1.06
Tomasi di Lampedusa	4.99 (0.02) 0.17	26.42 (1.43) 12.30	3.33 (0.14) 1.22	7.90 (0.14) 1.23
Verga	4.46 (0.05) 0.15	20.45 (0.78) 7.35	3.00 (0.10) 0.97	6.82 (0.06) 0.56
Global values	4.67 (0.006)	24.34 (0.29)	3.67 (0.04)	6.56 (0.03)

$$G_C = 46.7 \quad (3)$$

and the average value of G_F described by the regression line

$$G_F = 0.912 \times G - 37.1 \quad (4)$$

The correlation coefficient between G_F and G , Equation (4), is 0.932. The slope is 0.912, therefore, giving practically a 45° line. By considering the coefficient of variation, $100 \times 0.932^2 = 86.9\%$ of the data is explained by (4). **Figure 2(a)** shows also the average values of selected works listed in **Table 1** to locate them in this scatter plot. **Figure 2(b)** shows, superposed to the scattered values of P_p , the theoretical relationship between the average value of P_p , as a function of G , given, according to Equations (1a) and (3), by:

$$P_F = \frac{300}{G - 42.3} \quad (5)$$

The correlation between the experimental values of P_F and that calculated from (5) is 0.800. The correlation between the experimental values of P_F and G is -0.830 .

In conclusions, Equation (1a) can be rewritten by modifying the constant from 89 to 42.3, without significantly changing the numerical values of Equation (1), but now giving a meaning to the constant 42.3, as the minimum value G_{\min} , so that (1) can be written as:

$$G_{estim} = G_{\min} + G_F \quad (6)$$

From these results, it is evident that each author has his own “dynamics”, in the sense that each author modulates the length of sentences in a way significantly more ample than he does or, I should say, he could do with the length of words, and differently from other authors, as we can read in **Table 2**. We pass, for example, from 11.93 words per sentence (Cassola) to 44.27 words per sentence (Boccaccio), whereas the number of characters per word ranges only from 4.481 to 4.475, a much smaller range. Even if the two authors are spaced centuries apart, have very different literary style, write very different novels and address very different audiences—all characteristics well-known in the history of Italian Literature—both use words of very similar length.

The average number of characters per word, C_p , varies between 4.37 (Bembo, Sacchetti) and 5.01 (Salgari), a range equal to 0.64 characters per word which, compared to the global average value 4.67 (**Table 2**), corresponds to $\mp 6.8\%$ change. On the contrary, G_F varies from 6.94 (Boccaccio) to 25.65 (Cassola), with excursions in the range from -52% to $\mp 78\%$, compared to the global average value 14.40 (**Table 1**). Of course, the values of each text block can vary around the average, as for example **Figure 3** and **Figure 4** show for Boccaccio and Manzoni, because of different types of literary texts, such dialogues, descriptions, author’s considerations or comments, etc. On the other hand, the reader that wishes to read it all is exposed to the full variety of texts, which in any case must be read. In other words, what counts is the average value of a parameter,

not the variations that it can assume in each text block, as also Martin and Gottron underline [26].

By considering the above findings, we can state that G_C is practically a constant, $G_C = 46.70$, and that G can be approximated by (6).

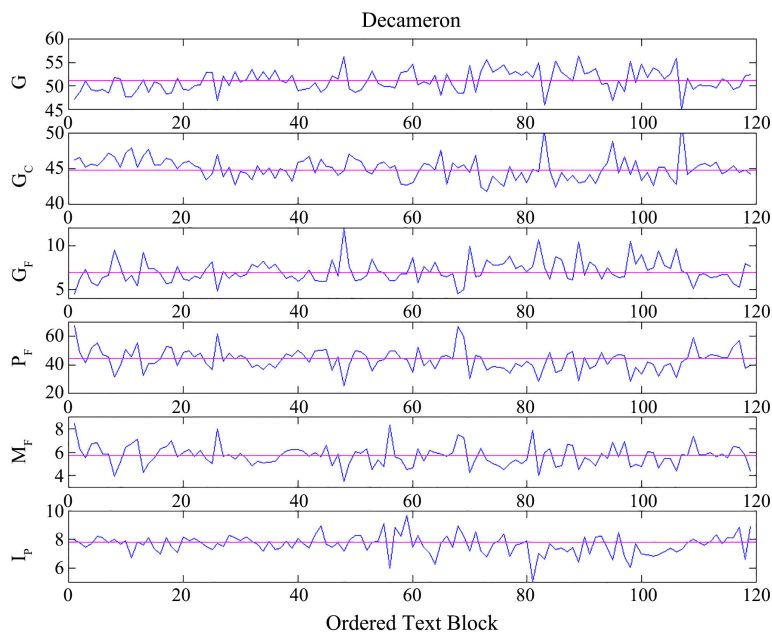


Figure 3. Ordered text-block series of G , G_C , G_F , P_F , M_F and I_p versus the ordered sequence of text blocks found in Boccaccio's *Decameron*. The text blocks are the novels told, on turn, by each character each day. The horizontal magenta lines give the average values (Table 1 and Table 2).

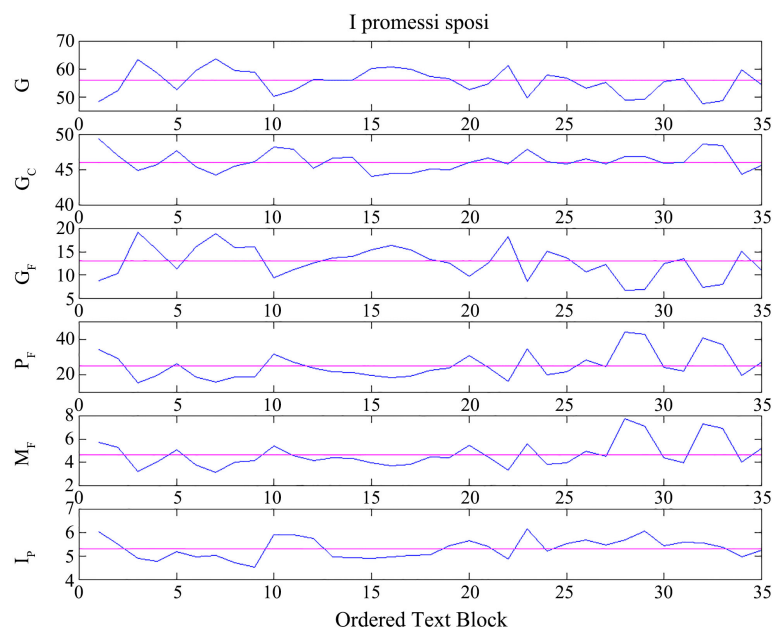


Figure 4. Ordered text-block series of G , G_C , G_F , P_F , M_F and I_p versus the ordered sequence of text blocks found in Manzoni's *I promessi sposi*. The text blocks are the chapters. The horizontal magenta lines give the average values (Table 1 and Table 2).

Figure 5(a) shows the scatter plot between the values calculated with Equation (6) by using the value of G_F of each text block, and the values calculated with Equation (1a), and the regression line between the two data sets. The slope is 0.998, in practice 1 (45° line), and the correlation coefficient is⁴ 0.932.

Defined the error $G - G_{estim}$, its average value is -0.1 , therefore 0 for any practical purpose, and its standard deviation is 2.14. For a constant readability level G , the latter value translates into an estimating error of school years required by at most 1 year, see **Figure 1**. **Figure 5(b)** shows that a normal (Gaussian) probability density function with zero average value and standard deviation 2.14 describes very well the error scattering.

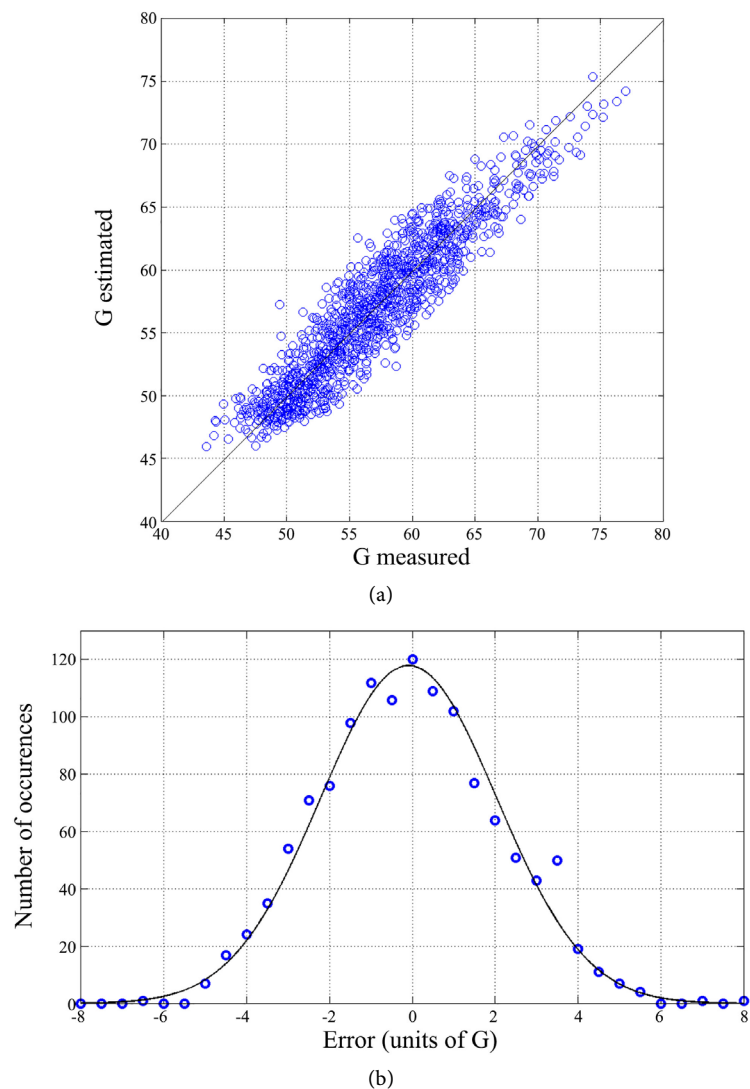


Figure 5. (a) Scatter plot of G_{estim} Equation (6), and the values G calculated with Equation (1). Also shown the regression line, in practice the 45° line $G = G_{estim}$; (b) Histogram of the error $G - G_{estim}$ (blue circles) and theoretical histogram (black line) due to a Gaussian (normal) density function with average value -0.1 and standard deviation 2.14.

⁴This value is the same as that of the couple (G, G_F) because G_F is linearly related to G .

Now, according to (6) it is obvious that the constant value G_{\min} can be set to zero, therefore making:

$$G \cong G_s = G_F \quad (7)$$

with the advantage that the scaled index G_s starts at 0. Now Equation (7) is not meant to be used to reduce any computability effort, as today Equation (1), like any other readability formula or other approaches, can be calculated by means of dedicated software, with no particular effort. Equation (7) is useful because underlines the fact that authors of the Italian Literature modulate much more the length of sentences, and each of them with personal style, than the length of words, and that the length of sentences substantially determines reading difficulty (as any Italian student knows when reading Boccaccio's *Decameron*, or Collodi's *Pinocchio!*), so that we could use **Figure 6**, as a guide, instead of **Figure 1**.

4. Characters, Words, Sentences, Punctuation Marks, Word and Time Intervals

Table 3 shows that, for any author, there is a large correlation, close to unity, between the number of characters and the number of words, as **Figure 7** directly shows. The correlation coefficient is 0.999 and the slope of the line $y = mx$ is $m = 4.67$ characters per word, equal to the average value (**Table 2**), because the correlation coefficient is very close to 1. On the average, every word in the Italian literature is made of 4.67 ± 0.006 characters, so that characters and words can be interchanged in any mathematical relationship.

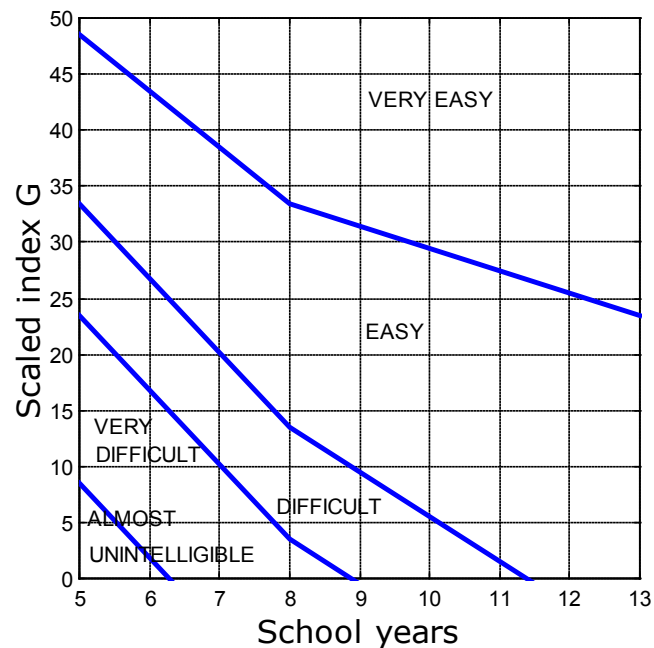


Figure 6. Scaled index G_s as a function of the number of school years attended in the Italian School System.

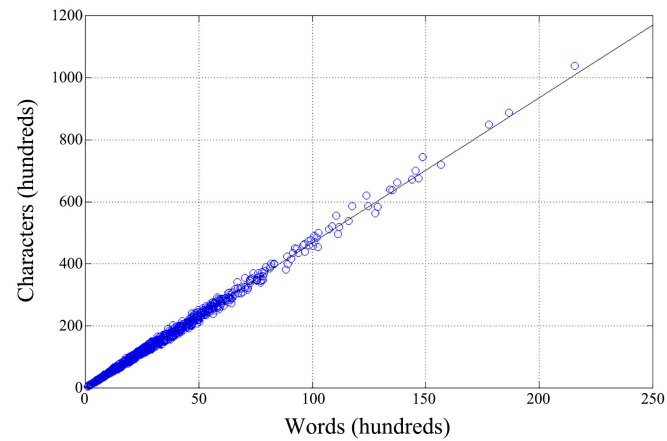


Figure 7. Scatter plot between the number of characters and the number of words (1260 text blocks), also shown the regression line (see **Table 3**).

Table 3. Correlation coefficient and slope (in parentheses) of the line $y = mx$ modelling the indicated variables.

Author	Characters vs. words	Words vs. sentences	Punctuation marks vs. sentences	Words vs. punctuation marks
Anonimo	0.999 (4.64)	0.951 (6.40)	0.989 (8.25)	0.969 (4.42)
Bembo	0.999 (4.31)	0.951 (31.03)	0.993 (5.99)	0.968 (5.21)
Boccaccio	0.999 (4.47)	0.975 (43.54)	0.993 (7.82)	0.980 (5.57)
Buzzati (D)	0.998 (5.12)	0.947 (16.98)	0.973 (6.60)	0.980 (2.57)
Buzzati (B)	0.999 (4.82)	0.976 (14.52)	0.985 (6.33)	0.993 (2.29)
Calvino (B)	0.997 (4.62)	0.805 (17.21)	0.956 (6.59)	0.918 (2.64)
Calvino (M)	0.999 (4.73)	0.909 (16.87)	0.980 (6.56)	0.937 (2.57)
Cassola	0.999 (4.48)	0.948 (11.38)	0.989 (5.56)	0.981 (2.06)
Collodi	0.998 (4.61)	0.881 (15.57)	0.986 (6.16)	0.925 (2.54)
Da Ponte	0.997 (4.71)	0.657 (4.41)	0.958 (6.89)	0.711 (3.55)
Deledda	0.998 (4.51)	0.813 (14.52)	0.915 (5.96)	0.937 (2.44)
D'Azeglio	0.999 (4.64)	0.820 (28.83)	0.983 (7.33)	0.892 (3.95)
De Amicis	0.999 (4.62)	0.978 (14.98)	0.997 (5.35)	0.987 (2.82)
De Marchi	0.999 (4.72)	0.987 (18.88)	0.998 (7.07)	0.990 (2.67)
D'Annunzio	0.999 (4.91)	0.908 (14.65)	0.959 (5.88)	0.969 (2.52)
Eco	0.999 (4.80)	0.945 (20.47)	0.990 (7.47)	0.961 (2.75)
Fogazzaro (S)	0.999 (4.79)	0.984 (14.58)	0.995 (6.33)	0.994 (2.31)
Fogazzaro (P)	0.999 (4.73)	0.970 (15.26)	0.987 (6.01)	0.976 (2.55)
Gadda	0.999 (4.77)	0.848 (17.39)	0.984 (4.98)	0.910 (3.51)
Grossi	0.998 (4.59)	0.696 (23.86)	0.879 (6.24)	0.932 (3.89)
Leopardi	0.997 (4.72)	0.754 (21.93)	0.970 (6.57)	0.844 (3.43)
Levi	0.997 (4.73)	0.811 (21.98)	0.986 (5.69)	0.809 (3.86)

Continued

Machiavelli	0.999 (4.73)	0.972 (40.01)	0.997 (6.44)	0.975 (6.21)
Manzoni (PS)	0.992 (4.60)	0.619 (21.87)	0.943 (5.25)	0.783 (4.20)
Manzoni (FL)	0.992 (4.75)	0.670 (28.12)	0.884 (6.96)	0.822 (4.05)
Moravia (I)	0.9998 (4.80)	0.970 (32.79)	0.996 (6.65)	0.979 (4.94)
Moravia (C)	0.997 (4.56)	0.914 (29.39)	0.939 (7.23)	0.940 (4.06)
Pavese (B)	0.995 (4.54)	0.563 (11.93)	0.852 (5.93)	0.845 (2.019)
Pavese (F)	0.996 (4.46)	0.697 (16.45)	0.883 (6.71)	0.798 (2.45)
Pellico	0.991 (4.80)	0.673 (16.29)	0.770 (6.34)	0.634 (2.55)
Pirandello	0.999 (4.63)	0.894 (13.75)	0.982 (4.86)	0.952 (2.84)
Sacchetti	0.996 (4.37)	0.754 (21.18)	0.918 (5.73)	0.912 (3.72)
Salernitano	0.999 (4.39)	0.888 (16.36)	0.991 (5.09)	0.938 (3.26)
Salgari (C)	0.998 (4.98)	0.910 (14.41)	0.970 (6.32)	0.968 (2.28)
Salgari (M)	0.996 (5.01)	0.670 (14.48)	0.901 (6.16)	0.840 (2.36)
Svevo	0.9998 (4.86)	0.973 (15.62)	0.989 (7.69)	0.971 (2.03)
Tomasi di Lampedusa	0.999 (5.00)	0.875 (25.18)	0.990 (7.88)	0.921 (3.21)
Verga	0.999 (4.45)	0.965 (19.46)	0.996 (6.81)	0.973 (2.86)
Global values	0.998 (4.68)	0.877 (18.61)	0.972 (6.25)	0.913 (2.99)

The relationship between words and sentences behaves differently. For each author a line $y = mx$ still describes, usually very well, their relationship (see **Table 2** and **Table 3**), but with different slope, as **Figure 8** shows. The average number of words per sentence varies from 11.93 (Cassola) to 44.47 (Boccaccio) and these values affect very much the sentence term G_p , which varies from 25.65 (Cassola) to 6.94 (Boccaccio). In **Figure 8**, we can notice that there is an angular range where all authors fall, a range that has collapsed into a line in **Figure 7** because of a very tight, and equal for all authors, relationship between characters and words. Moreover, notice that the value of p/f calculated from the average G_p , i.e. $p/f = 300/G_p$, is always smaller or at most equal⁵ to the average value of the ratio p/f (**Table 2**).

Defined the total number of punctuation marks (sum of commas, semicolons, colons, question marks, exclamation marks, ellipsis, periods) contained in a text, **Figure 9** shows the scatter plot between this value and the number of sentences for each text block. Once more, for any author the relationship is a line $y = mx$ with correlation coefficients close to 1 (**Table 3**), but with different slopes, the latter close to the average number of punctuation marks per sentence. For example, in Boccaccio, the average number of punctuation marks per sentence is $M_F = 5.69$ (**Table 2**), whereas the slope⁶ of the corresponding line is $m = 5.57$ (**Table 3**).

⁵It can be proved, with Cauchy-Schwarz inequality, that the average value of $1/x$ ($x = p/f = 300/G_p$), is always less or equal to the reciprocal of the average value of x .

⁶The slope $m = y/x$ has dimensions of words per punctuation mark, like the word interval I_p .

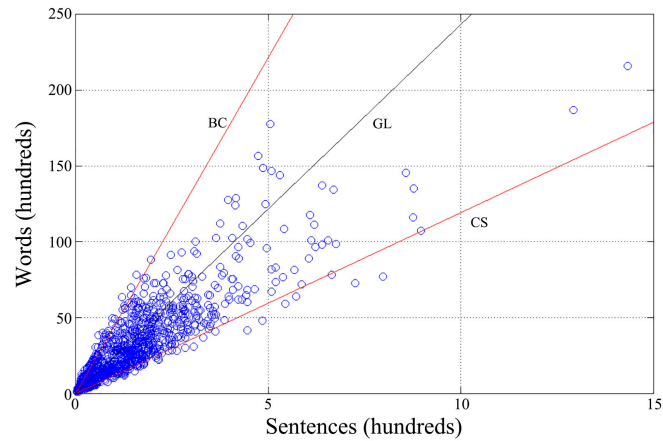


Figure 8. Scatter plot between the number of words and the number of sentences (1260 text blocks). *BC* refers to Boccaccio, *CS* refers to Cassola, *GL* refers to the global values. The two authors represent approximate bounds to the angular region.

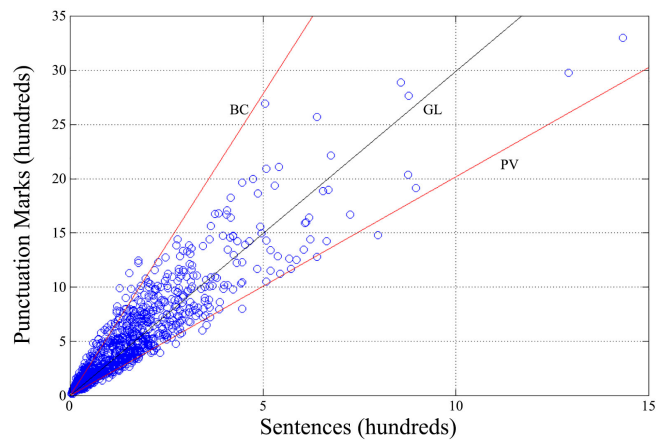


Figure 9. Scatter plot between the number of punctuation marks and the number of sentences (1260 text blocks). *BC* refers to Boccaccio, *PV* refers to Pavese (*La bella estate*), *GL* refers to the global values. The two authors represent approximate bounds to the angular region. The ratio between the ordinate and the abscissa gives the word interval.

An interesting comparison among different authors and their literary works can be done by considering the number of words per punctuation mark, that is to say, the average number of words between two successive punctuation marks, a random variable that is the *word interval* I_p mentioned before, defined by:

$$I_p = \frac{p}{i} \quad (8)$$

The word interval I_p is very robust against changing habits in the use of punctuation marks throughout decades. Punctuation marks are used for two goals: 1) improving readability by making lexical and sentence constituents of texts more easily recognizable, 2) introducing pause [27], and the two goals can coincide [28] [29]. In the last decades, in Italian, there has been a reduced use of semicolons in favor of periods [30], but this change does not affect I_p but only the number of words per sentence.

The values of I_p listed in **Table 2** vary from 5.64 (Cassola) to 7.8 (Boccaccio). For any author, the linear model $y = mx$ is still valid, as the high correlation coefficients listed in **Table 3** and **Figure 10** show. The slopes of the lines are very close to the averages, namely 5.56 and 7.82 respectively, because of correlation coefficients⁷ close to 1.

Finally, **Figure 11** shows the scatter plot between G , G_O , G_F and I_p . We can notice that G_F (and G) is significantly correlated with I_p through an inverse proportionality. This result is very interesting because it links the readability of a text, the index G , or G_F , to I_p , another author's distinctive characteristic. Moreover, the word interval has other very interesting and intriguing relationships, as section 5 shows.

5. Comparing Different Literary Texts: Distances

A large number of texts produced today in several forms, both in hard copies and digital formats, such as books, journals, technical reports and others, have prompted several methods for fast automatic information retrieval, document classification, including authorship attribution. The approach is to represent documents with n -grams using vector representation of particular text features [31]. In this model, the similarity between two documents is estimated using the cosine of the angle between the corresponding vectors. This approach depends mainly on the similarity of the vocabulary used in the texts, while the characters and syntax are ignored. A more complex approach represents textual data in more detail [31]. These new techniques, implemented with complex software, are useful when, together with other tasks, automatic authorship attribution and verification are required.

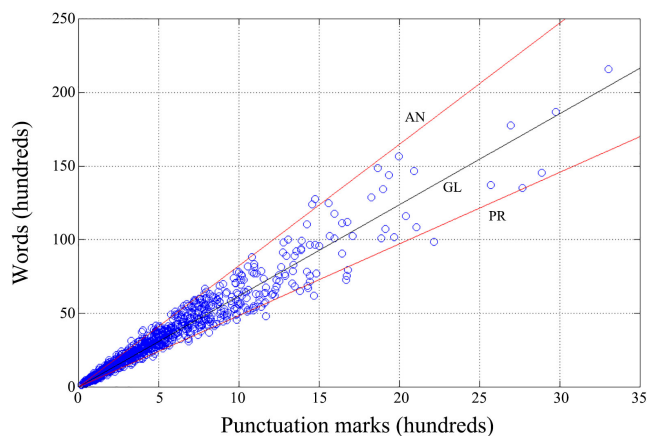


Figure 10. Scatter plot between the number of words and the number of punctuation marks (1260 text blocks). *AN* refers to Anonymus, *PR* refers to Pirandello, *GL* refers to the global values. The two authors represent approximate bounds to the angular region. The ratio between the ordinate and the abscissa gives the word interval.

⁷The ratio between P_F (column 3 of **Table 2**) and M_F (column 4) is another estimate of the word interval I_p (column 5). The value so calculated and that of column 5 almost coincide because the correlation coefficient is close to 1. In other words, the ratio of the averages (column 3 divided by column 4) is practically equal to the average value of the ratio (column 5).

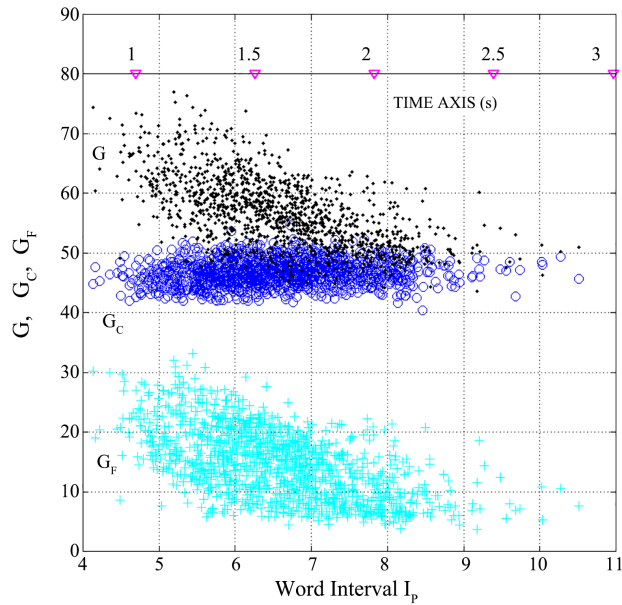


Figure 11. Scatter plot between G (black dots), G_c (blue circles), G_f (cyan crosses) and the word interval I_p . The top time axis refers to the time interval I_T (Section 6).

In the case of the literary texts considered in this paper, it is more interesting to compare the statistical characteristics of different authors or different texts of the same author, by using the data reported in **Tables 1-3**, instead of using the more complex methods reviewed by [32]. For this purpose, the parameters that are most significant are the four random variables defined before: C_p , P_p , M_f and I_p , because they represent fundamental indices and are mostly uncorrelated, except the couple (M_p, I_p) , as **Table 4** shows. These parameters are suitable to assess similarities and differences of texts much better, as I show next, than the cosine of the angle between any two vectors. Therefore, in this section, I define absolute and relative “distances” of texts by considering the following six vectors of components⁸ x and y : $\mathbf{R}_1 = (C_p, P_f)$, $\mathbf{R}_2 = (M_f, P_f)$, $\mathbf{R}_3 = (I_p, P_f)$, $\mathbf{R}_4 = (C_p, M_f)$, $\mathbf{R}_5 = (I_p, M_f)$, $\mathbf{R}_6 = (I_p, C_p)$.

Now, considering the six vectors just defined, the average cosine similarity S between two documents (literary texts) D_1 and D_2 can be computed as:

$$S(D_1, D_2) = \frac{1}{6} \sum_{k=1}^6 \cos(\mathbf{R}_{D_{1k}}, \mathbf{R}_{D_{2k}}) \tag{9}$$

where $\cos(\mathbf{R}_{D_{1k}}, \mathbf{R}_{D_{2k}})$ is the cosine of the angle formed by the two vectors $\mathbf{R}_{D_{1k}}, \mathbf{R}_{D_{2k}}$. If all pairs of vectors were collinear (aligned), then $\cos(\mathbf{R}_{D_{1k}}, \mathbf{R}_{D_{2k}}) = 1$, the similarity would be maximum, $S = 1$. If all pairs of vectors were orthogonal $\cos(\mathbf{R}_{D_{1k}}, \mathbf{R}_{D_{2k}}) = 0$, the similarity would be zero, $S = 0$. According to this criterion, two collinear vectors of very different length (the magnitude of the vector) will be classified as identical because $S_k = 1$, a conclusion that cannot be ac-

⁸The choice of which parameter represents the component x or y is not important. Once the choice is made, the numerical results will depend on it, but not the relative comparisons and general conclusions

cepted. This is a serious drawback of the cosine similarity.

Figure 12 shows the scatter plot between the average value of S , calculated by considering all text blocks, and the readability index G . Any text block is compared also to another text block of the same literary text (but not with itself). The choice of not excluding the other text blocks of the same literary text leads to a simple and straight software code, which, however, does not affect the general conclusion arrived at by observing the scatter plot shown in **Figure 12**: there is no correlation between S and G , therefore S does not meaningfully discriminate between any two texts when the angle formed by their vectors is close to zero.

Table 4. Linear correlation coefficients between the indicated pairs of random variables (1260 text blocks).

	G	G_C	G_F	C_P	P_F	M_F	I_P
G	1	-0.215	0.932	-0.215	-0.830	-0.769	-0.607
G_C	-0.215	1	0.154	1	-0.163	-0.223	0.121
G_F	0.932	0.154	1	0.154	-0.900	-0.854	-0.569
C_P	-0.215	1	0.154	1	-0.163	-0.228	0.121
P_F	-0.830	-0.163	-0.900	-0.163	1	0.248	0.594
M_F	-0.769	-0.223	-0.854	-0.228	0.248	1	0.937
I_P	-0.607	0.121	-0.569	0.121	0.594	0.937	1

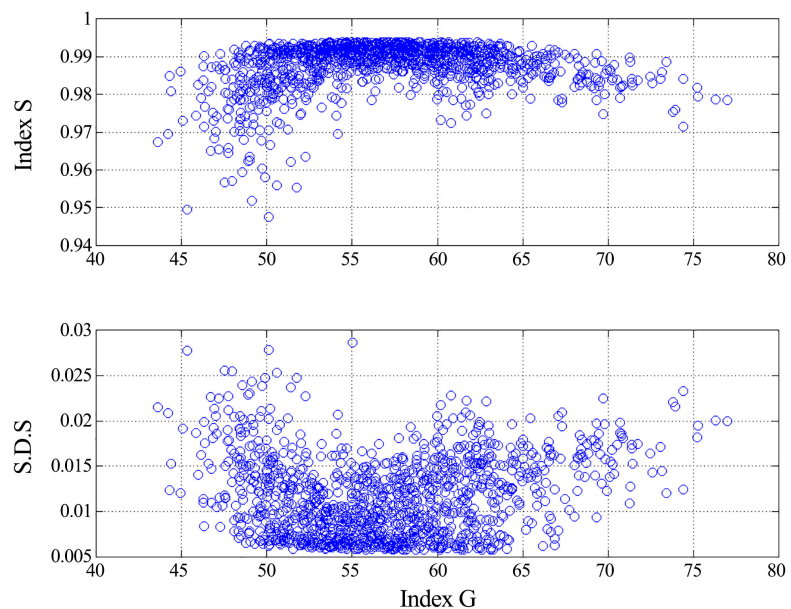


Figure 12. Upper panel: Scatter plot between the average similarity index S of a text block, out of 1260 in total, about all others, and the corresponding readability index G . Lower panel: standard deviation. The total amount of data used to calculate average and standard deviation is given by $1260 \times (1260 - 1) = 1586340$.

Now, a better choice for comparing literary texts is to consider the “distance” of any text block from the origin of x and y axes⁹, given by the magnitude of the resulting vector R :

$$R = |\mathbf{R}| = \left| \sum_{k=1}^6 \mathbf{R}_k \right| \quad (10)$$

With this vectorial representation, a text block ends up in a point of coordinates x and y in the first Cartesian quadrant, as **Figure 13** shows. The end point of the vectors with components given by the average values of the literary texts (obtainable from **Tables 1-3**) is also shown.

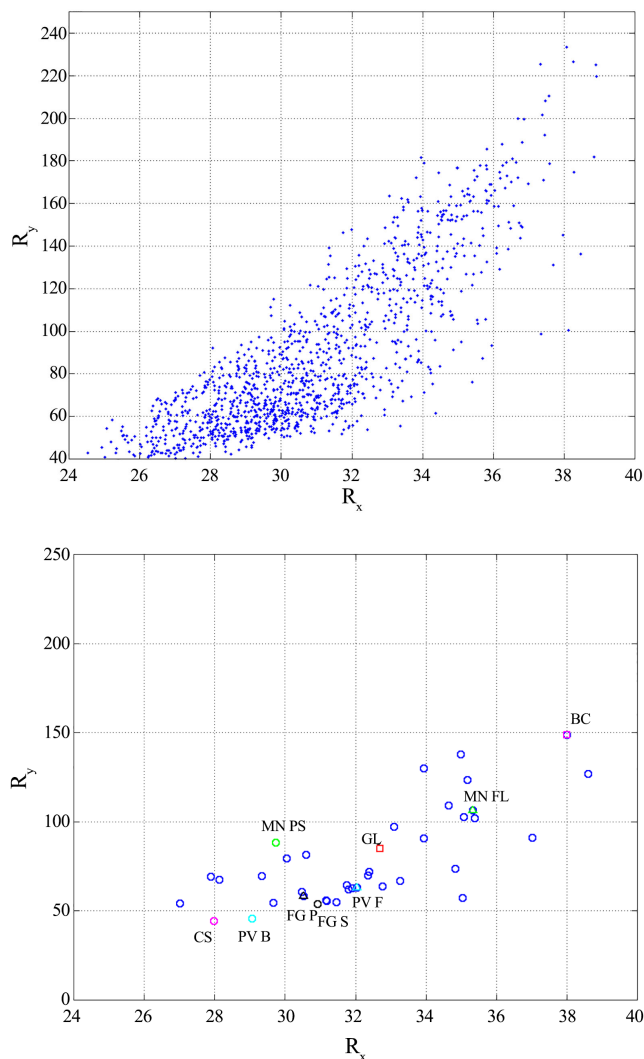


Figure 13. Scatter plot between the two components of the distance R for all 1260 text block (upper panel), and that calculated from the average values shown in **Tables 1-3** (lower panel). *CS* = Cassola, *PV B* = Pavese *La bella estate*, *PV F* = Pavese *La luna e i falò*, *MN PS* = Manzoni *I promessi sposi*, *MN FL* = Manzoni *Fermo e Lucia*, *FG* = Fogazzaro *Il santo* and *Piccolo mondo antico*, *BC* = Boccaccio, *GL* = global values (“barycentre”).

⁹From vector analysis, the two components of a vector are given by $x = \sum_{k=1}^6 x_k$, $y = \sum_{k=1}^6 y_k$. The magnitude is given by the Euclidean (Pythagorean) distance $R = \sqrt{x^2 + y^2}$.

It is very interesting, for example, to compare the vector representing Manzoni's masterpiece¹⁰ *I promessi sposi* (published in 1840) and that representing Manzoni's *Fermo e Lucia* (published in 1827). The latter novel was the first version of *I promessi sposi* and the great improvement pursued by Manzoni in many years of revision, well known to scholars of Italian literature, is also observable mathematically: the absolute distances are 113.9 for *Fermo e Lucia* and 93.7 for *I promessi sposi*, the relative distance is 20.3, a significant fraction of the entire range spanning from Cassola to Boccaccio, whose relative distance is 106.6¹¹. Other interesting observations are the coincidence of the two vectors representing the novels by Fogazzaro, and the difference between the novels by Pavese, etc. With this tool, the scholars of Italian literature (even if not accustomed to using mathematics in their research) could find some objective confirmation of their literary studies concerning an author, as exemplified in the case of Manzoni.

The efficacy of R can be appreciated in **Figure 14**, which shows the scatter plot between R and G , and between its angle $\varphi = \arctan\left(\frac{y}{x}\right)$, expressed in degrees, and G . The black lines describe very well the relationships between them, given by:

$$R = 21.16 \times \frac{G}{G - 42.3} \quad (11a)$$

$$\varphi = -0.875 \times G + 118.3 \quad (11b)$$

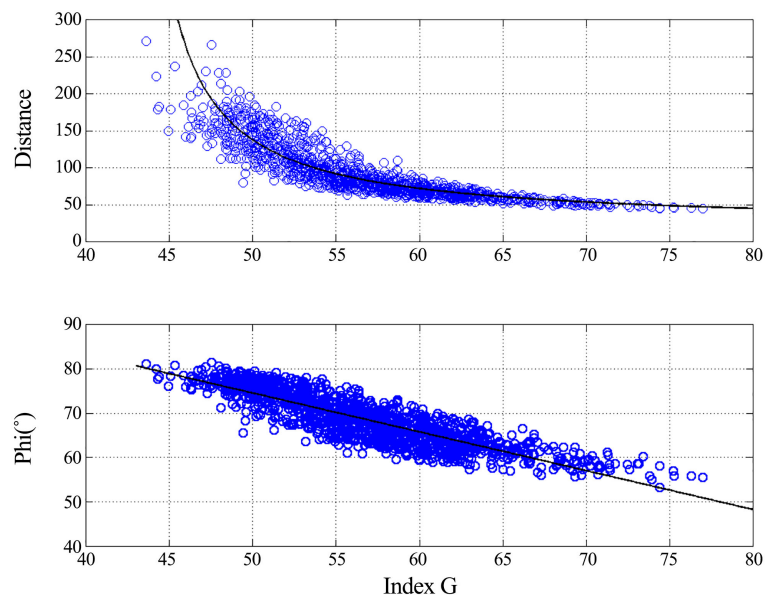


Figure 14. Scatter plot between R and G (upper panel) and between $\varphi = \arctan\left(\frac{y}{x}\right)$ and G (lower panel).

¹⁰A compulsory reading in any Italian High School.

¹¹Notice that distances are distorted, if measured on the graph of Figure 13, because the abscissa (x scale) is expanded compared to the ordinate (y scale).

The correlation coefficient is -0.832 , for the couple (R, G) and -0.867 for the couple (φ, G) . The correlation coefficient between measured and estimated values of R through (11a) is 0.802 between the measured and estimated values of φ with (11b) is¹² 0.867 . In conclusion, the magnitude (distance) R and the angle φ of the vector \mathbf{R} are very well correlated with the readability index G .

6. Word Interval, Miller's 7 ± 2 Law and Short-Term Memory Capacity

The range of the word interval I_p , shown in **Figure 11**, is very similar to the range mentioned in Miller's law 7 ± 2 , although the short-term memory capacity of data for which chunking is restricted is 4 ± 1 [33] [34] [35] [36] [37]. For words, namely data that can be restricted (*i.e.*, "compressed") by chunking, it seems that the average value is not 7 but around 5 to 6 [21], almost the average value of the word interval 6.56 (**Table 2**). Now, as the range from 5 to 9 in Miller's law corresponds to 95% of the occurrences [37], it is correct to compare Miller's interval with the dispersion of the word interval in single text block shown in **Figure 11**, where we can see values ranging from 4 to 10.5, close to Miller's law range.

The probability density function and the complementary probability distribution of I_p are shown in **Figure 15**. From the lower panel we can see that 95% of the samples (probabilities between 0.025 and 0.975) fall in the range from 4.6 to 8.6, which coincides, in practice, with Miller's range 7 ± 2 . The most likely value (the mode of the distribution) is 6.3 and the median is 6.5. The experimental density can be modelled with a log-normal model with three parameters:

$$f(I_p) = \frac{1}{\sqrt{2\pi}\sigma_{I_p}(I_p-1)} \exp\left\{-\frac{1}{2}\left[\frac{\log(I_p-1)-\mu_{I_p}}{\sigma_{I_p}}\right]^2\right\}, \quad I_p \geq 1 \quad (12a)$$

with confidence level in excess of 99.99% (chi-square test) [38]. The log-normal probability density is valid only for $I_p \geq 1$ being $I_p = 1$ the minimum theoretical value of this variable (a single sentence made of only 1 word).

The theoretical constants of Equation (12) are obtained as follows. Given the average value $m_{I_p} = 6.56$ and the standard deviation $s_{I_p} = 1.01$, of the random variable I_p for the 1260 text blocks, the standard deviation σ_{I_p} and the average value μ_{I_p} of the random variable $\log(I_p)$ of a three-parameter log-normal probability density function [39] are given (natural logs) by:

$$\sigma_{I_p}^2 = \log\left[\left(\frac{s_{I_p}}{m_{I_p-1}}\right)^2 + 1\right] = 0.0326 \quad (12b)$$

$$\mu_{I_p} = \log\left[\left(m_{I_p} - 1\right) - \frac{\sigma_{I_p}^2}{2}\right] = 1.698 \quad (12c)$$

¹²This value is the same as that of the couple (R, G) because the two parameters are related by the linear relationship (11b).

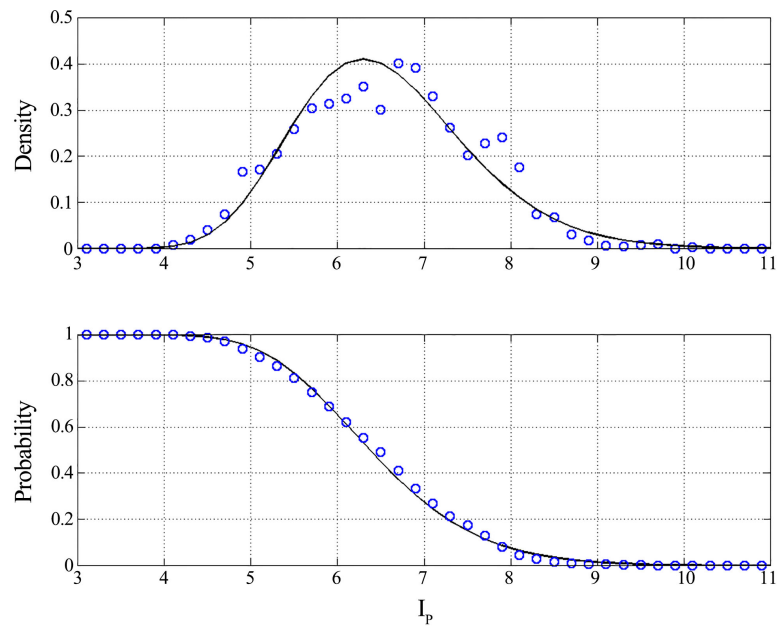


Figure 15. Probability density function (upper panel, blue circles) and the complementary probability distribution (lower panel, blue circles) of I_p for 1260 text blocks. The lower panel shows the probability that the value reported in abscissa (x axis) is exceeded. The black continuous lines are the theoretical density and distribution of a three-parameter log-normal model [39].

The theoretical mode (the most likely value) is given by $\mu_o = \exp(\mu_{I_p} - \sigma_{I_p}^2) + 1 = 6.297$. Notice, however, that the experimental mode observed in **Figure 15** is near 7, very close to the central value of Miller's range and that an underlying mixture of three shifted probability distributions may be a better model. However, for the purpose of linking the word interval I_p to the short-term memory capacity this analysis may be sufficient.

These results may be explained, at least empirically, according to the way our mind is thought to memorize “chunks” of information in the short-term memory. When we start reading a sentence, our mind tries to predict its full meaning from what has been read up to that point, as it seems that can be concluded from the experiments of Jarvella [40]. Only when a punctuation mark is found, our mind can better understand the meaning of the text. The longer and more twisted is the sentence, the longer the ideas remain deferred until the mind can establish the meaning of the sentence from all its words. In this case, the text is less readable, a result quantitatively expressed by the empirical Equation (1a) for Italian.

Figure 16 shows the scatter plot between I_p and P_F for all the text blocks, together with the non-linear regression line (best-fit line) that models, on the average, I_p versus P_F for the Italian Literature, given by:

$$I_p = (I_{P_{\infty}} - 1) \times \left[1 - e^{-\frac{(P_F - 1)}{(P_{F_0} - 1)}} \right] + 1 \quad (13)$$

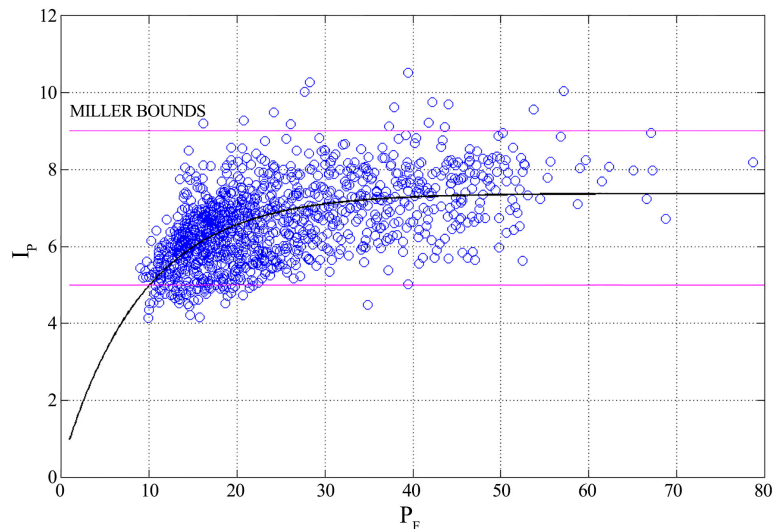


Figure 16. Scatter plot between the word interval I_p and the number of words per sentence P_F for all text blocks. The continuous black line refers to the best fit given by Equation (13). Also shown the 7 ∓ 2 bounds of Miller's Law (magenta lines).

where $I_{p_\infty} = 7.37$ words per inter-punctuations is the horizontal asymptote, and $P_{F_0} = 10.22$ words per sentence is the value of P_F at which the exponential in Equation (13) falls at $1/e$ of its maximum value. Notice that: 1) most data fall in the Miller's range 7 ∓ 2 ; 2) the asymptotic value (7.37) is very close to the center value of Miller's range; 3) $P_F = P_{F_0} = 10.22$ is the abscissa corresponding to the lower bound of Miller's law.

These results can be explained as follows. As the number of words increases, the number of word intervals can increase but not linearly, because the short-term memory cannot hold, approximately, a number of words larger than that empirically described by Miller's Law. In other words, scatter plots like that shown in **Figure 16** drawn for other Literatures—classical Greek and Latin Literatures, or modern languages Literatures—should give us an insight into the short-term memory capacity required to their readers.

In conclusion, the range of the word interval is similar to Miller's law range. The values found for each author seems to set the size of the short-term memory capacity that their readers should have to read the literary work more easily. For example, the reader of Boccaccio's *Decameron* should have a short-term memory able to memorize $I_p = 7.79 \mp 0.06$ chunks, on the average, whereas the reader of Collodi's *Pinocchio* needs only a memory of capacity $I_p = 6.19 \mp 0.08$ chunks. Now, if our conjecture is found reliable after more studies concerning short-term memory and brain, the link between G_F and G through Equation (6), would appear justified and natural.

The word interval can be translated into a *time interval* if we consider the average reading speed of Italian, estimated in 188 words per minute [41]. In this case, the average time interval corresponding to the word interval, expressed in seconds, is given by:

$$I_T = 60 \times \frac{I_P}{188} \quad (14)$$

The time axis drawn in **Figure 11** is useful to convert I_p into I_T . The values of I_p shown in the scatter plot, now read as time interval, according to the time scale agree very well with the intervals of time so that the immediate memory records the stimulus for later memorizing it in the short term memory, ranging from 1 to about 2-3 seconds [10] [13] [22] [23] [42] [43].

The results, relating I_T and I_p to fundamental and accessible characteristics of short-term memory, are very interesting and should be furtherly pursued by experts. Moreover, the same studies can be done on ancient languages, such as Greek and Latin, to test the expected capacity and response time of the short-term memory of these ancient and well-educated readers, partially already done for the Greek of the New Testament [24].

7. Technical and Scientific Writings

Technical and scientific writings (papers, essays, etc.) ask more to their readers. A preliminary investigation was done on short scientific texts published in the Italian popular science magazines *Le Scienze* and *Sapere* (because today is rare to find original scientific papers written in Italian), in a popular scientific book and newspaper editorials gave the results listed in **Table 5**. In this analysis, mathematical expressions, tables, legends have not been considered. From **Table 5** we can notice some clear differences from the results of novels: words are on the average longer, the readability index G is lower, the word interval is longer. These results are not surprising because technical and scientific writings use long technical words, deal with abstract meaning with articulated and elaborated sentences resulting in long sentences with series of subordinate clauses. Of course, the reader of these texts expects to find technical and abstract terms of his field, or specialty, and would not understand the text if these elements were absent.

8. Conclusions and Future Developments

Statistics of languages have been calculated for several western languages, mostly by counting characters, words, sentences, word rankings. Some of these parameters are also the main “ingredients” of classical readability formulae. Revisiting the readability formula of Italian, known with the acronym GULPEASE, shows that of the two terms that determine the readability index G —the semantic index G_C proportional to the number of characters per word, and the syntactic index G_P proportional to the reciprocal of the number of words per sentence— G_P is dominant because G_C is, in practice, constant for any author. From these results, it is evident that each author modulates the length of sentences more freely than what he can do with word length and in different ways from author to author.

Table 5. Statistics of some recent texts extracted from popular scientific literature and daily newspapers comments and short essays.

Texts	c	p	f	G	G_C	G_F	C_p	P_F	M_F	I_p
Bellone ¹³	188,536	36,122	1505	49.31 (0.53)	52.19 (0.36)	12.50 (0.25)	5.22 (0.04)	24.23 (0.52)	2.86 (0.06)	8.51 (0.15)
Popular scientific articles ¹⁴	188,537	33,797	1241	44.23 (0.63)	55.79 (0.37)	11.02 (0.37)	5.58 (0.04)	27.97 (0.90)	2.95 (0.11)	9.57 (0.24)
Newspapers editorials ¹⁵	60,281	11,919	479	50.48 (0.59)	50.58 (0.47)	12.06 (0.45)	5.06 (0.05)	25.28 (0.88)	2.99 (0.16)	8.67 (0.44)

For any author, any couple of text variables can be described by a linear relationship $y = mx$ but with different slope m from author to author, except for the relationship between characters and words, which is unique.

The most important relationship I have found is that between the short-term memory capacity, described by Miller's "7 \pm 2 law", and what I have termed the *word interval*, a new random variable defined as the average number of words between two successive punctuation marks. The word interval can be converted into a *time interval* through the average reading speed. The *word interval* is numerically spread in a range very alike to that found in Miller's law, and the *time interval* is spread in a range very alike to that found in the studies on short-term memory response time. The connection between the word interval (and time interval) and short-term memory appears, at least empirically, justified and natural.

For ancient languages, no longer spoken by a people, but rich in literary texts that have founded the Western civilization, such as Greek or Latin, nobody can make reliable experiments, as those reported in the references recalled above. These ancient languages, however, have left us a huge library of literary and (few) scientific texts. Besides the traditional count of characters, words and sentences, the study of their word interval statistics should bring us a flavor of the short-term memory features of these ancient readers, and this can be done very easily, as I have done for Italian. A preliminary analysis of a large number of Greek and Latin literary texts shows results very similar to those reported in this paper, therefore evidencing some universal and long-lasting characteristics of western languages and their readers.

In conclusion, it seems that there is a possible direct and interesting connection between readability formulae and reader's capacity of short-term memory capacity and response time. As short-term memory features can be related to other cognitive parameters [20], this relationship seems to be very useful. However, its relationship with Miller's law should be further investigated because the word interval is another parameter that can be used to design a text, together with readability formulae, for better matching expected reader's characteristics.

¹³Bellone, E. (1999) *Spazio e tempo nella nuova scienza*. Carocci, 136 pages.

¹⁴*Le Scienze, Scienze e ricerche*, 2017 issues.

¹⁵*Il Corriere della Sera, La Repubblica, Il Sole 24 ore*, 2018.

Acknowledgments

I am very grateful to *Liber Liber* (<https://www.liberliber.it/online/opere/libri/>) for providing the digital versions of the classics of Italian Literature. Without this large texts bank, I could not have ever obtained the results reported in the paper.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

- [1] Grzybeck, P. (2007) History and Methodology of Word Length Studies. In: *Contributions to the Science of Text and Language*, Springer, Dordrecht, 15-90. https://doi.org/10.1007/1-4020-4068-7_2
- [2] DuBay, W.H. (2004) *The Principles of Readability*. Impact Information, Costa Mesa.
- [3] DuBay, W.H. (2006) *The Classic Readability Studies*. Impact Information, Costa Mesa.
- [4] Anderson, P.V. (1991) *Technical Writing: A Reader-Centered Approach*. 2nd Edition, Harcourt Brace Jovanovich, Fort Worth.
- [5] Burnett, R.E. (2004) *Technical Communication*. 3rd Edition, Wadsworth Publishing Company, Belmont.
- [6] Matricciani, E. (2007) *La scrittura tecnico-scientifica*. Casa Editrice Ambrosiana, Milano.
- [7] Saaty, T.L. and Ozdemir, M.S. (2003) Why the Magic Number Seven Plus or Minus Two. *Mathematical and Computer Modelling*, **38**, 233-244. [https://doi.org/10.1016/S0895-7177\(03\)90083-5](https://doi.org/10.1016/S0895-7177(03)90083-5)
- [8] Zamanian, M. and Heydari, P. (2012) Readability of Texts: State of the Art. *Theory and Practice in Language Studies*, **2**, 43-53. <https://doi.org/10.4304/tpls.2.1.43-53>
- [9] Collins-Thompson, K. (2014) Computational Assessment of Text Readability: A Survey of Past, in Present and Future Research, Recent Advances in Automatic Readability Assessment and Text Simplification, ITL. *International Journal of Applied Linguistics*, **165**, 97-135. <https://doi.org/10.1075/itl.165.2.01col>
- [10] Baddeley, A.D., Thomson, N. and Buchanan, M. (1975) Word Length and the Structure of Short-Term Memory. *Journal of Verbal Learning and Verbal Behavior*, **14**, 575-589. [https://doi.org/10.1016/S0022-5371\(75\)80045-4](https://doi.org/10.1016/S0022-5371(75)80045-4)
- [11] Barrouillet, P. and Camos, V. (2012) As Time Goes By: Temporal Constraints in Working Memory. *Current Directions in Psychological Science*, **21**, 413-419. <https://doi.org/10.1177/0963721412459513>
- [12] Jones, G. and Macken, B. (2015) Questioning Short-Term Memory and Its Measurements: Why Digit Span Measures Long-Term Associative Learning. *Cognition*, **144**, 1-13. <https://doi.org/10.1016/j.cognition.2015.07.009>
- [13] Chekaf, M., Cowan, N. and Mathy, F. (2016) Chunk Formation in Immediate Memory and How It Relates to Data Compression. *Cognition*, **155**, 96-107. <https://doi.org/10.1016/j.cognition.2016.05.024>
- [14] Bailin, A. and Graftstein, A. (2001) The Linguistic Assumptions Underlying Readability

- bility Formulae: A Critique. *Language & Communication*, **21**, 285-301. [https://doi.org/10.1016/S0271-5309\(01\)00005-2](https://doi.org/10.1016/S0271-5309(01)00005-2)
- [15] Benjamin, R.G. (2012) Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty. *Educational Psychology Review*, **24**, 63-88. <https://doi.org/10.1007/s10648-011-9181-8>
- [16] Vajjala, S., Meurers, D., Eitel, A. and Scheiter, K. (2016) Towards Grounding Computational Linguistic Approaches to Readability: Modelling Reader-Text Interaction for Easy and Difficult Texts. *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*, Osaka, 11-17 December 2016, 38-48.
- [17] Dell'Orletta, F., Montemagni, S. and Venturi, G. (2011) Read-It: Assessing Readability of Italian Texts with a View to Text Simplification. *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*, Edinburgh, 30 July 2011, 73-83.
- [18] De Mauro, T. (1980) Guida all'uso delle parole. Editori Riuniti, Roma.
- [19] Atvars, A. (2016) Eye Movement Analyses for Obtaining Readability Formula for Latvian Texts for Primary School. *Procedia Computer Science*, **104**, 477-484. <https://doi.org/10.1016/j.procs.2017.01.162>
- [20] Conway, A.R.A., Cowan, N., Bunting, M.F., Theriault, D.J. and Minkoff, S.R.B. (2002) A Latent Variable Analysis of Working Memory Capacity, Short-Term Memory Capacity, Processing Speed, and General Fluid Intelligence. *Intelligence*, **30**, 163-183. [https://doi.org/10.1016/S0160-2896\(01\)00096-4](https://doi.org/10.1016/S0160-2896(01)00096-4)
- [21] Miller, G.A. (1955) The Magical Number Seven, plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychological Review*, **101**, 343-352. <https://doi.org/10.1037/0033-295X.101.2.343>
- [22] Grondin, S. (2000) A Temporal Account of the Limited Processing Capacity. *Behavioral and Brain Sciences*, **24**, 122-123. <https://doi.org/10.1017/S0140525X01303928>
- [23] Muter, P. (2000) The Nature of Forgetting from Short-Term Memory. *Behavioral and Brain Sciences*, **24**, 134. <https://doi.org/10.1017/S0140525X01423922>
- [24] Matricciani, E. and De Caro, L. (2019) A Deep-Language Mathematical Analysis of Gospels, Acts and Revelation. *Religions*, **10**, 257. <https://doi.org/10.3390/rel10040257>
- [25] Lucisano, P. and Piemontese, M.E. (1988) GULPEASE: Una formula per la predizione della difficoltà dei testi in lingua italiana. *Scuola e città*, **3**, 110-124.
- [26] Martin, L. and Gottron, T. (2012) Readability and the Web. *Future Internet*, **4**, 238-252. <https://doi.org/10.3390/fi4010238>
- [27] Parkes, M.B. (2016) Pause and Effect. An Introduction to the History of Punctuation in the West. Routledge, Abingdon-on-Thames, 343 p. <https://doi.org/10.4324/9781315247243>
- [28] Maraschio, N. (1993) Grafia e ortografia: Evoluzione e codificazione. In: L. Serianni, & P. Trifone (Eds.), *Storia della lingua italiana: I luoghi della codificazione*, Einaudi, Torino, 139-227.
- [29] Mortara Garavelli, B. (2003) Prontuario di punteggiatura, Editori Laterza.
- [30] Serianni, L. (2001) Sul punto e virgola nell'italiano contemporaneo. *Studi Linguistici italiani*, **27**, 248-255.
- [31] Gómez-Adorno, E., Sidorov, G., Pinto, D., Vilariño, D. and Gelbukh, A. (2016) Automatic Authorship Detection Using Textual Patterns Extracted from Integrated Syntactic Graphs. *Sensors*, **16**, 1374. <https://doi.org/10.3390/s16091374>
- [32] Stamatatos, E. (2009) A Survey of Modern Authorship Attribution Methods. *Jour-*

-
- nal of the American Society for Information Science and Technology*, **60**, 538-556.
<https://doi.org/10.1002/asi.21001>
- [33] Cowan, N. (2000) The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity. *Behavioral and Brain Sciences*, **24**, 87-114.
<https://doi.org/10.1017/S0140525X01003922>
- [34] Bachelder, B.-L. (2001) The Magical Number 4 = 7: Span Theory on Capacity Limitations. *Behavioral and Brain Sciences*, **24**, 116-117.
<https://doi.org/10.1017/S0140525X01243921>
- [35] Chen, Z. and Cowan, N. (2005) Chunk Limits and Length Limits in Immediate Recall: A Reconciliation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **3**, 1235-1249. <https://doi.org/10.1037/0278-7393.31.6.1235>
- [36] Mathy, F. and Feldman, J. (2012) What's Magic about Magic Numbers? Chunking and Data Compression in Short-Term Memory. *Cognition*, **122**, 346-362.
<https://doi.org/10.1016/j.cognition.2011.11.003>
- [37] Gignac, G.E. (2015) The Magical Numbers 7 and 4 Are Resistant to the Flynn Effect: No Evidence for Increases in Forward or Backward Recall across 85 Years of Data. *Intelligence*, **48**, 85-95. <https://doi.org/10.1016/j.intell.2014.11.001>
- [38] Papoulis, A. (1990) Probability & Statistics. Prentice Hall, Upper Saddle River.
- [39] Bury, K.V. (1975) Statistical Models in Applied Science. John Wiley, Hoboken.
- [40] Jarvella, R.J. (1971) Syntactic Processing of Connected Speech. *Journal of Verbal Learning and Verbal Behavior*, **10**, 409-416.
[https://doi.org/10.1016/S0022-5371\(71\)80040-3](https://doi.org/10.1016/S0022-5371(71)80040-3)
- [41] Trauzettel-Klosinski, S. and Dietz, K. (2012) Standardized Assessment of Reading Performance: The New International Reading Speed Texts IReST. *Investigative Ophthalmology & Visual Science*, **53**, 5452-5461.
<https://doi.org/10.1167/iovs.11-8284>
- [42] Mandler, G. and Shebo, B.J. (1982) Subitizing: An Analysis of Its Component Processes. *Journal of Experimental Psychology: General*, **111**, 1-22.
<https://doi.org/10.1037//0096-3445.111.1.1>
- [43] Pothos, E.M. and Jola, P. (2000) Linguistic Structure and Short-Term Memory. *Behavioral and Brain Sciences*, **24**, 138-139.
<https://doi.org/10.1017/S0140525X01463928>

Appendix: List of mathematical symbols with their meaning.

Symbol	Meaning
C_p	characters per word
G	GULPEASE readability index
G_c	semantic index
G_f	syntactic index
G_{\min}	minimum readability index
G_s	scaled readability index
G_{estim}	estimated readability index
I_T	time interval corresponding to I_p
I_p	words per punctuation marks (word interval)
M_F	punctuation marks per sentence
P_F	words per sentence
c	characters
f	sentences
m	slope of line
m_{I_p}	average value of I_p
n	number of text blocks
p	words
p_T	total words
S_{I_p}	standard deviation of I_p
v	average square value
μ	average value
μ_{I_p}	average value of $\log(I_p)$
μ_o	mode of $\log(I_p)$
σ	standard deviation relative to text blocks
σ_{I_p}	standard deviation of $\log(I_p)$
σ_r	standard deviation relative to 1000-word text blocks
σ_μ	standard deviation of average value