 doi:10.20944/preprints201811.0505.v1

Article

Italian Throughout Seven Centuries of Literature: Deep Language Statistics And Their Relationship With Miller's 7 ∓ 2 Law and Short–Term Memory

Emilio Matricciani

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano
Correspondence: Emilio.Matricciani@polimi.it; Tel.: +39–2399–3639

Abstract: Statistics of languages are 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, in author's opinion, that between the short–term memory capacity, described by Miller's "7 ∓ 2 law", 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* is spread in the same of 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, and should further investigated. Technical and scientific writings (papers, essays etc.) ask more to their readers. A preliminary investigation of these texts shows clear differences: 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 Greek or Latin, could bring us a flavor of the short term–memory features of these ancient readers.

Keywords: Italian, readability, GULPEASE, literature, statistics, characters, words, sentences, punctuation marks, short–term memory, word interval, time interval.

---

1. Introduction

Statistics of languages have been calculated for several western languages, mostly by counting characters, words, sentences, word rankings (Grzybeck, 2007). Some of these parameters are also the main "ingredients" of classical readability formulae. First developed in the United States (DuBay, 2004), readability formulae are applicable to any language, once the mathematical expression for that

 doi:10.20944/preprints201811.0505.v1

particular language is developed and assessed experimentally (DuBay, 2006). 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 (Anderson, 1991) (Burnett, 1994) (Matricciani, 2007), 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 (DuBay, 2006) , (Saaty and Ozdemir, 2003), (Zamanian and Heydari, 2012), (Collins and Thompson, 2014), 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 behaviour (Baddeley et al., 1975), (Barrouillest and Camos, 2012), (Jones and Macken, 2015), (Chekaf et al., 2016). In this paper, I show that a statistical relationship can be found for Italian, by examining a large number of literary texts written since XIV century, thus revitalizing, in my opinion, the classical readability formula approach.

The classical readability formulae have been, in fact, criticized because they focus on a limited set of superficial text features, rough approximations of the linguistic factors at play in readability assessment (Bailin and Graftstein, 2001). 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" 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, *relative* readability, or 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 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 (Benjamin, 2012), (Vajjala et al., 2016), to capture more complex linguistic features (Dell'Orletta et al. 2011). Benjamin (Benjamin, 2012), 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 1980's (De Mauro, 1980).

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 (Vajjala, 2016), (Atvars, 2016). For Italian, the work by Dell'Orletta and colleagues (Dell'Orletta et al. 2011) 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.

 doi:10.20944/preprints201811.0505.v1

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 give 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, in my opinion, 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 behaviour, 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 a 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 (e.g. Dell'Orletta et al., 2011). 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 (Conway et al., 2002), a task beyond the scope of this paper and author's expertise.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2018 doi:10.20944/preprints201811.0505.v1

The most important relationship I have found is, in my opinion, that between the short−term memory capacity, described by Miller's "7 ∓2 law" (Miller, 1955), 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, 2015), and the *time interval* is spread in a range very alike to that found in the studies on short−term memory response time (Baddeley et al., 1975) (Grondin, 2000,) (Muter, 2000). The connection between the word interval (and time interval) and short−term memory appears, at least empirically, justified and natural.

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 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. These results will be reported next.

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

a) The impact of semantic and syntactic indices on the readability index (all defined in Section 2)
b) The relationship of these indices with the newly defined "word interval" and "time interval"
c) The "distance", absolute and relative, of literary texts by defining meaningful vectors based on characters, words, sentences, punctuation marks.
d) 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 discusses the "distance" of literay texts, Section 5 introduces word and the time intervals and their empirical relationships with short−term memory features, and finally Section 6 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 (Lucisano and Piemontese, 1988), is given by:

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2018 doi:10.20944/preprints201811.0505.v1

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, as shown by (Lucisano and Piemontese, 1988) and summarized in Figure 1. The larger G, the more readable the text is. In (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 number of sentences contained in the p words (a list of mathematical symbols is reported in the Appendix). Defined the terms:

$$G_C = 10 \times \frac{c}{p} \qquad (2a)$$

$$G_F = 300 \times \frac{f}{p} \qquad (2b)$$

equation (1a) is written as:

$$G = 89 - G_C + G_F \qquad (1b)$$

We analyze first equation (1), by means of standard statistics of its addends, because, as other readability formulae, it contains 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, reading difficulty is due to syntax and therefore we term 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, difficulty is due to semantics, and therefore we term 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.

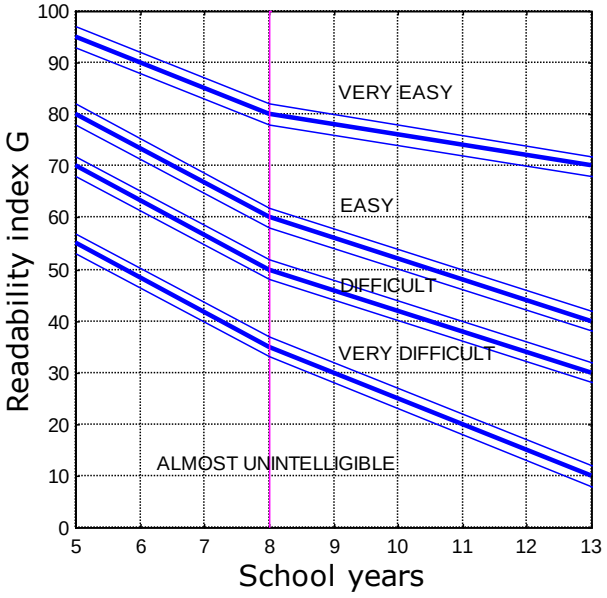

Figure 1. Readability index G of Italian, as a function of the number of school years attended (in Italy high school lasts 5 years, kids attend it up to 19 years old). Elaborated from (Lucisano and Piemontese, 1988). The

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2018 doi:10.20944/preprints201811.0505.v1

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).

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 equation (1) to classical literary works of a large number of Italian writers[1], 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 edition[2].

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 (eq. (2a)), and the syntactic index G_F, given by the reciprocal of the number of words per sentence, multiplied by 300 (eq. (2b)), affect very differently the final value of G (eq. (1b)). Table 1 lists the average values of G, G_C e 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[3]. From the results reported in Table 1, it is evident that G_C changes much less than G_F, a feature highlighted in the scatter plot of Figure 2a, which shows G_C and G_F versus G, for each text block (1260 in total, with different number of words) found in the listed literary works.

---

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

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

[3] Let n be the number of text blocks contained in a literay 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 tex block $i-th$. For the standard deviation of the average value, we calculate first the average square value $v = 300^2 \sum_{i=1}^{i=n} \left(\frac{f_i}{p_i}\right)^2 \times \frac{p_i}{p_T}$ and the standard deviation in the n text blocks $\sigma = \sqrt{v - \mu^2}$, and finally we calculate $\sigma_\mu = \frac{\sigma}{\sqrt{n}}$. In this way different literay works can be realibly compared with regard to the avergae value of any parameter, regardless of the choice of the length of text blocks.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2018 doi:10.20944/preprints201811.0505.v1

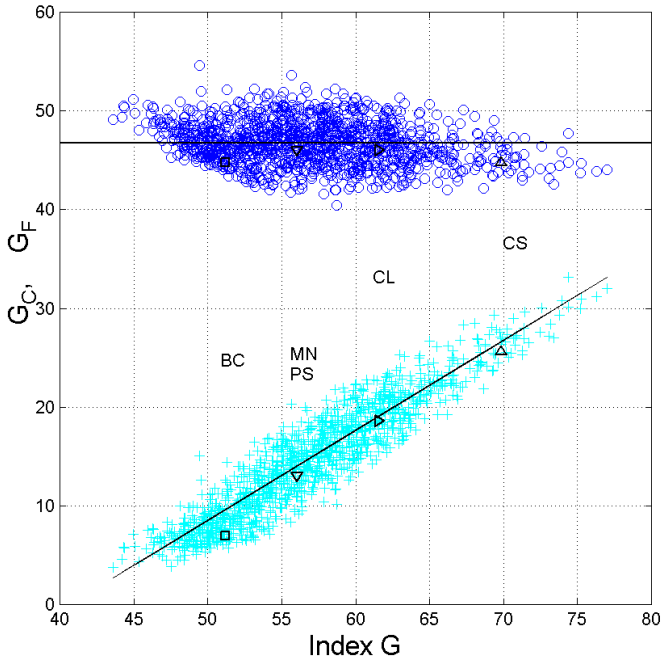

Figure 2a: G_C (blue circles) and G_F (cyan crosses) versus G, for all 1260 text blocks. BC=Boccaccio, MN PS=Manzoni (*I promessi sposi*), CL=Collodi, CS=Cassola. The horizontal line is the average value of G_C, equation (3), the slant line is the average value of G_F, equation (4).

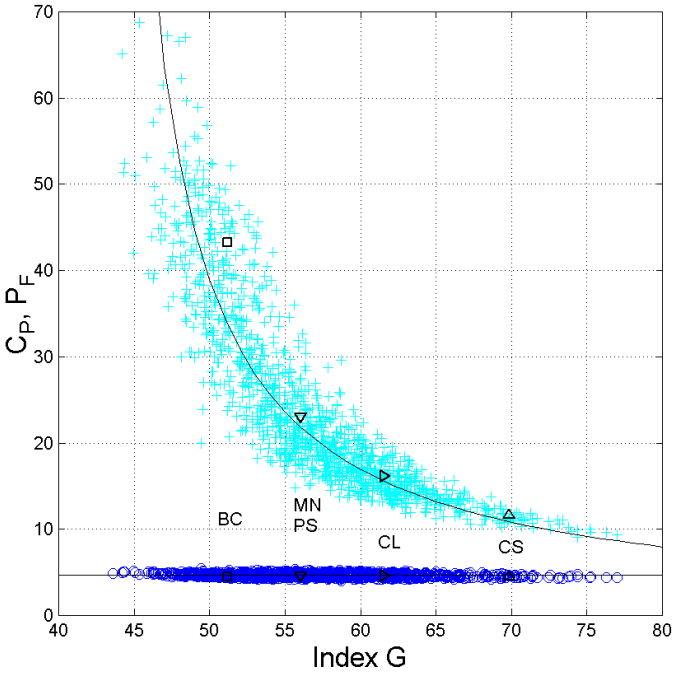

Figura 2b: Scatter plots of $C_P = c/p$ vs. G (blu dots) and $P_F = p/f$ vs. G (cyan crosses). BC=Boccaccio, MN PS=Manzoni (*I promessi sposi*), CL=Collodi, CS=Cassola. The black continuous lines are given by (3) and (5).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2018 doi:10.20944/preprints201811.0505.v1

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 words[4]. 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 (see also Figure xx).

Author, literary work, century	Characters	Words	Sentences	G	G_C	G_F
Anonymous (*I Fioretti di San Francesco*, XIV)	180056	38681	1064	50.70 (0.30) 1.84	46.55 (0.163 1.01	8.25 (0.24) 1.48
Bembo Pietro (*Prose*, XV-XVI)	295614	67572	1925	53.80 0.66) 5.42	43.75 (0.30) 2.44	8.55 (0.44) 3.65
Boccaccio Giovanni (*Decameron*, XIV)	1190417	266033	6147	51.18 (0.17) 2.85	44.75 (0.11) 1.84	6.94 (0.10) 1.63
Buzzati Dino (*Il deserto dei tartari*, XX)	292974	57402	3311	55.27 (0.54) 4.08	51.04 (0.27) 2.03	17.30 (0.51) 3.86
Buzzati Dino (*La boutique del mistero*, XX)	302894	62771	4219	60.91 (0.81) 6.40	48.25 (0.19) 1.50	20.16 (0.69) 5.44
Calvino (*Il barone rampante*, XX)	330420	71340	3864	58.93 (0.84) 7.10	46.32 (0.23) 1.91	16.25 (0.77) 6.53
Calvino Italo (*Marcovaldo*, XX)	161952	34206	2000	59.19 (0.85) 4.99	47.35 (0.24) 1.42	17.54 (0.74) 4.30
Cassola Carlo (*La ragazza di Bube*, XX)	307819	68698	5873	69.84 (1.11) 9.22	44.81 (0.22) 1.84	25.65 (0.96) 7.95
Collodi Carlo (*Pinocchio*, XIX)	186849	40642	2512	61.57 (0.79) 5.04	45.97 (0.22) 1.41	18.54 (0.70) 4.47
Da Ponte Lorenzo (*Vita*, XVIII-XIX)	646024	137054	5459	53.81 (0.50)	47.14 (0.14)	11.95 (0.43)

[4] The 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 $n_r = p_T/p_r$ and then scaling σ as $\sigma_r = \sigma \times \sqrt{\frac{n_r}{n}} = \sigma_\mu \sqrt{n_r}$.

 doi:10.20944/preprints201811.0505.v1

				5.87	1.65	5.04
Deledda Grazia (*Canne al vento*, XX, Nobel Prize 1926)	276552	61375	4184	64.39	45.06	20.45
				(0.92)	(0.18)	(0.79)
				7.21	1.39	6.22
D'Azeglio Massimo (*Ettore Fieramosca*, XIX)	424259	91464	3182	53.05	46.39	10.44
				(0.54)	(0.17)	(0.45)
				5.15	1.62	4.33
De Amicis Edmondo (*Cuore*, XIX)	376792	82770	4775	60.78	45.52	17.31
				(0.55)	(0.15)	(0.54)
				5.01	1.39	4.94
De Marchi Emilio (*Demetrio Panelli*, XIX)	471451	100328	5363	58.05	46.99	16.04
				0.44	0.14	0.35
				4.37	1.43	3.48
D'Annunzio Gabriele (*Le novelle delle Pescara*, XX)	244055	49688	3027	58.16	49.12	18.28
				(0.81)	(0.16)	(0.77)
				5.74	1.12	5.44
Eco Umberto (*Il nome della rosa*, XX)	821707	170676	8490	55.78	48.14	14.92
				(0.52)	(0.11)	(0.49)
				6.82	1.47	6.35
Fogazzaro (*Il santo*, XIX-XX)	467990	97616	6637	61.46	47.94	20.40
				(0.72)	(0.10)	(0.65)
				7.16	1.02	6.40
Fogazzaro (*Piccolo mondo antico*, XIX-XX)	528659	112706	7069	61.46	47.94	20.40
				(0.57)	(0.16)	(0.48)
				6.06	1.65	5.12
Gadda (*Quer pasticciaccio brutto*… XX)	474359	99631	5596	58.24	47.61	16.85
				(1.08)	(0.24)	(1.00)
				10.77	2.35	9.99
Grossi Tommaso (*Marco Visconti*, XIX)	637311	138900	5301	54.57	45.88	11.45
				0.63	0.12	0.55
				7.39	1.42	6.53
Leopardi Giacomo (*Operette morali*, XIX)	322991	68699	2694	53.76	47.01	11.77
				(1.17)	(0.32)	(0.96)
				8.46	2.62	6.46
Levi (*Cristo si è fermato a Eboli*)	383893	81092	3611	55.02	47.34	13.36
				(0.39)	(0.13)	(0.36)
				3.54	1.13	3.27
Machiavelli Niccolò (*Il principe*, XV-XVI)	130274	27680	702	49.54	47.06	7.61
				(0.33)	(0.21)	(0.21)
				1.75	1.10	1.09
Manzoni Alessandro (*I promessi sposi*, XIX)	1036728	225392	9766	56.00	46.00	13.00
				(0.69)	(0.21)	(0.52)
				10.29	3.18	7.76

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2018 doi:10.20944/preprints201811.0505.v1

Manzoni Alessandro (*Fermo e Lucia*, XIX)	1044997	219993	7496	51.72 (0.53) 2.84	47.50 (0.19) 2.84	10.22 (0.37) 5.54
Moravia Alberto (*Gli indifferenti*, XX)	471574	98084	2830	49.58 (0.40) 1.48	48.08 (0.15) 1.48	8.66 (0.40) 4.00
Moravia Alberto (*La ciociara*, XX)	577176	126550	4271	53.52 (0.38) 4.27	45.61 (0.24) 2.70	10.12 (0.33) 3.68
Pavese Cesare (*La bella estate*, XX)	116360	25650	2121	68.44 (0.90) 4.54	45.36 (0.17) 0.88	24.81 (0.85) 4.29
Pavese Cesare (*La luna e i falò*, XX)	194032	43442	2544	61.90 (0.65) 4.27	44.66 (0.14) 0.93	17.57 (0.66) 4.36
Pellico Silvio (*Le mie prigioni*, XIX)	252915	52644	3148	58.90 (0.37) 2.65	48.04 (0.12) 0.87	17.94 (0.32) 2.30
Pirandello Luigi (*Il fu Mattia Pascal*, Nobel Prize 1934)	345301	74544	5284	63.94 (1.01) 8.71	46.32 (0.20) 1.69	21.26 (0.93) 8.02
Sacchetti Franco (*Trecentonovelle*, XIV)	767538	175452	8060	59.04 (0.40) 5.35	43.75 (0.12) 1.62	13.78 (0.35) 4.63
Salernitano Masuccio (*Il Novellino*, XV)	152345	34623	1965	62.03 (0.89) 5.25	44.00 (0.20) 1.18	17.03 (0.91) 5.36
Salgari Emilio (*Il corsaro nero*, XIX-XX)	493213	98945	6686	59.42 (0.51) 5.09	49.85 (0.12) 1.21	20.27 (0.46) 4.58
Salgari Emilio (*I minatori dell'Alaska*, XIX-XX)	453614	90486	6094	59.07 (0.55) 5.22	50.13 (0.12) 1.14	20.20 (0.54) 5.11
Svevo Italo (*Senilità*, XX)	325221	66912	4236	59.39 (0.65) 5.33	48.60 (0.10) 0.81	19.00 (0.58) 4.78
Tomasi di Lampedusa (*Il gattopardo*, XX)	371853	74462	2893	50.72 (0.81) 6.96	49.94 (0.20) 1.73	11.66 (0.71) 6.09
Verga (*I Malavoglia*, XIX-XX)	393902	88277	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**	**46.70**	**14.40**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2018 doi:10.20944/preprints201811.0505.v1

		(0.16)	**(0.06)**	**(0.15)**

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 direct 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), features confirmed in Figure 2b, which shows the scatter plots of C_P vs. G and P_F vs. G In other words, the readability of a text using (1) is practically due only to the syntactic index G_F, therefore to the number of words per sentence. The two lines drawn in Figure 2a are given by the average value of G_C (Table 2):

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

and by the regression line

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

The correlation coefficient between G_F and G equation (4) is 0.932 and the slope 0.912 gives 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 2a shows also the average values of selected works listed in Table 1 to locate them in this scatter plot.

The theoretical range of G can be calculated by considering the theoretical range of G_F. The maximum value of G_F is found when P_F 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, 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). Figure 2b shows also, superposed to the scatterd values of P_F, the theoretical relationship between the average value of P_F, as a function of G, given, according to (1) and (3), by:

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

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 (1) 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 itself, as the minimum value G_{min}, so that (1) can be written as:

$$G = G_{min} + G_F \qquad (6)$$

From these results, it is evident that each author has his own "dynamics", in the sense that each modulates the length of sentences in a way significantly more ample than he does or, I should say, he can 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2018 doi:10.20944/preprints201811.0505.v1

Even if the two authors are spaced centuries apart, have very different 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 -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 Figures 3, 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, in my opinion, 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 (Martin and Gottron, 2012).

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

Figure 5a 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 (1), and the regression line between the two data sets. The slope is 0.998, in practice 1 (45° line), and the correlation coefficient is[5] 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 5b shows that a normal (Gaussian) probability density function with zero average value and standard deviation 2.14 describes very well the error scattering.

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

$$G_s = G_F \qquad (7)$$

with the advantage that the scaled index G_s starts at 0. Now (7) is not meant to be used to reduce any computability effort, as today equation (1), as any other readability formula or other approaches, can be calculated by means of dedicated software, with no particluar effort. In our opinion (7) is useful because inderlines 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 guide, instead of Figure (1).

[5] This value is the same of that of the couple (G, G_F) because G_F is linearly related to G.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2018 doi:10.20944/preprints201811.0505.v1

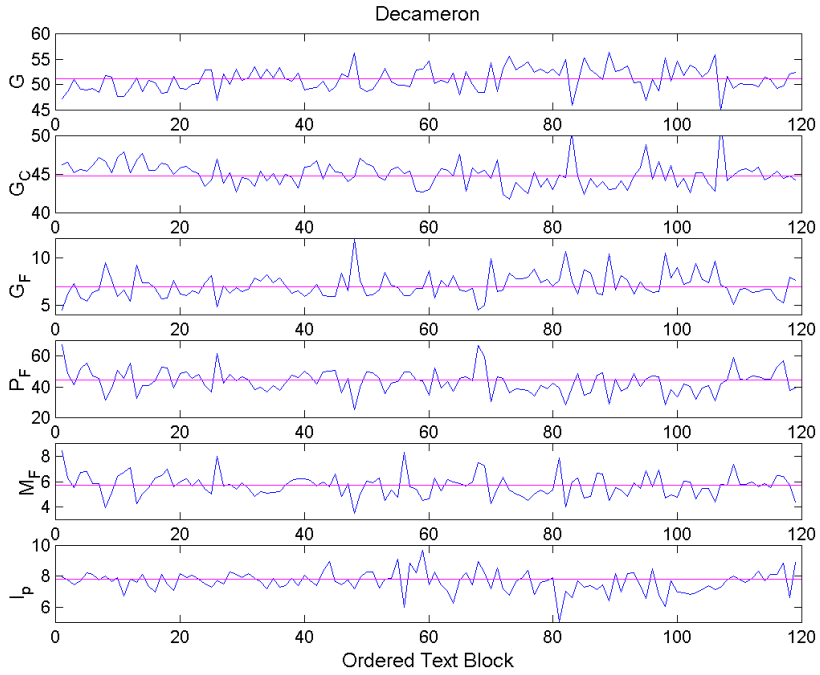

Figure 3: Ordered text–block series of G, G_C, G_F, P_F, M_F and I_p, versus the ordered sequence of text blocks found in Boccaccio's *Decameron*. The text blocks are the novels told, on turn, by each character each day. The horizontal magenta lines give the average values (Tables 1, 2)

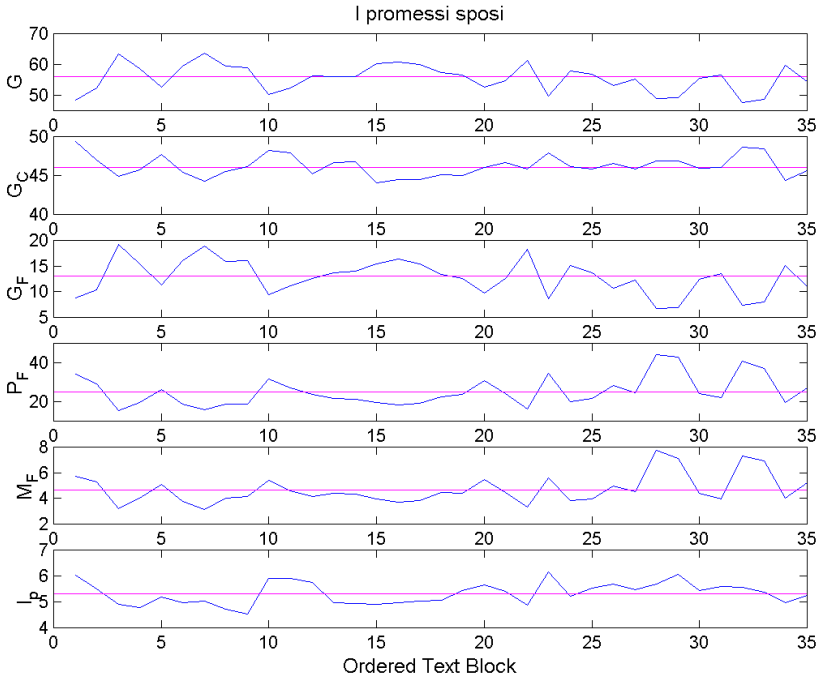

Figure 4: Ordered text–block series series of G, G_C, G_F, P_F, M_F and I_p, versus the ordered sequence of text blocks found in Manzoni's *I promessi sposi*. The text blocks are the chapters. The horizontal magenta lines give the average values (Tables 1, 2)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2018 doi:10.20944/preprints201811.0505.v1

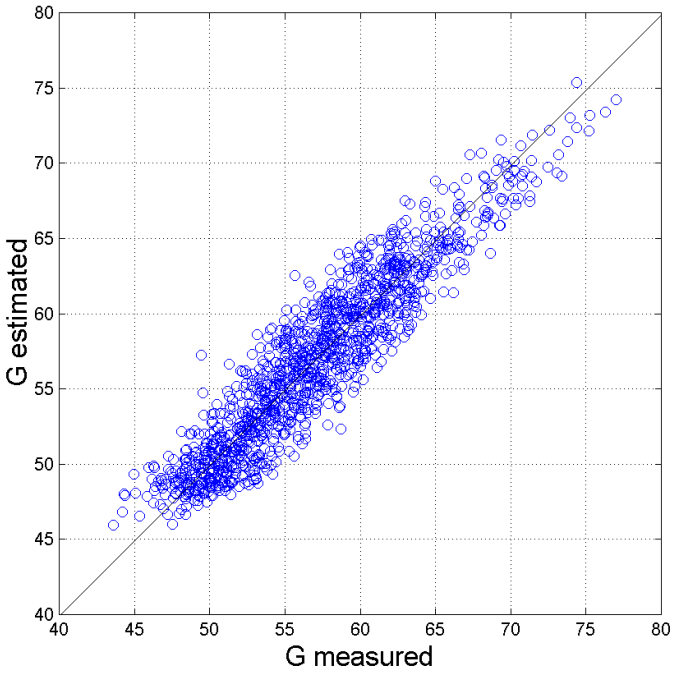

Figure 5a: Scatter plot of G_{estim} equation (6), and the values G calculated with equation (1). Also shown the regression line, in practice the 45° line $G_{estim} = G$.

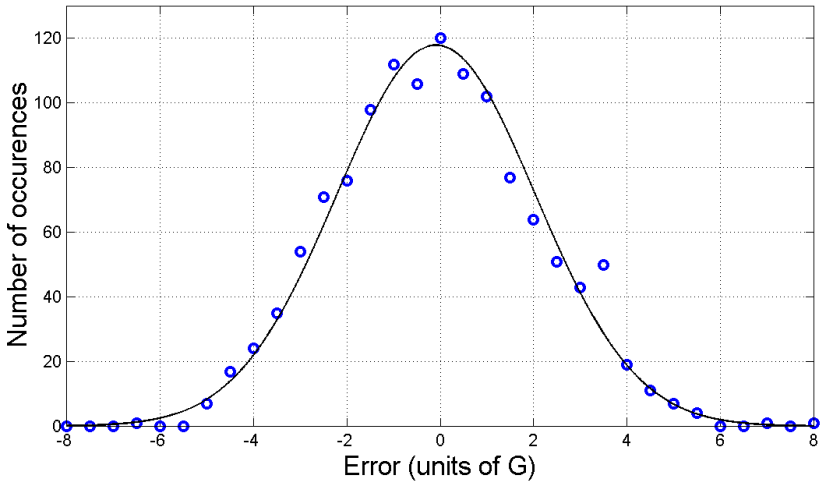

Figure 5b: Histogram of the error $G - G_{estim}$ (blue circles) and theoretical histogram (black line) due to a Gaussian (normal) density function with average value -0.1 and standard deviation 2.14.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2018 doi:10.20944/preprints201811.0505.v1

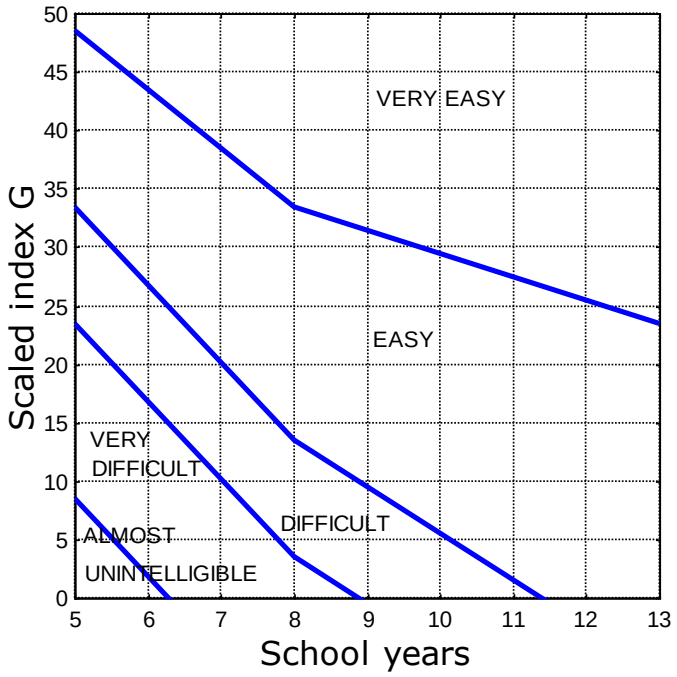

Figure 6: Scaled index G_S as a function of the number of school years attended.

4. Characters, words, sentences, punctuation marks and word interval

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.

The relationship between words and sentences behaves differently. For each author a line $y = mx$ still describes, usually very well, their relationship (see Tables 2 and 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 syntactic term G_F, 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_F, i.e. $p/f = 300/G_F$, is always smaller or at most equal[6] 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

[6] It can be proved, with Cauchy–Schwarz inequality, that the average value of $1/x$ ($x = p/f = 300/G_F$), is always less or equal to the reciprocal of the averge value of x.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2018 doi:10.20944/preprints201811.0505.v1

Boccaccio the average number of punctuation marks per sentence is M_F =5.69 (Table 2), whereas the slope[7] of the corresponding line is $m = 5.57$ (Table 3).

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:

$$I_P = \frac{p}{i} \qquad (8)$$

This parameter is very robust against changing habits in the use of punctuation marks throughout decades. Punctuation marks are used for two goals: i) improving readability by making lexical and syntactic constituents of texts more easily recognizable, ii) introducing pause (Parkes, 1992), and the two goals can coincide (Maraschio, 1993), (Mortara Garavelli, 2003). In the last decades, in Italian there has been a reduced use of semicolons in favour of periods (Serianni, 2001), but this change does not affect I_P but only the number of words per sentence.

The values of I_P listed 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 coefficients[8] close to 1.

Table 2: Average values of 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)	37.70 (1.19)	4.56 (0.12)	8.24 (0.11)
	0.10	7.39	0.74	0.67
Bembo	4.37 (0.03)	37.91 (2.16)	5.92 (0.25)	6.42 (0.34)
	0.24	17.72	1.98	1.02
Boccaccio	4.48 (0.01)	44.27 (0.59)	5.69 (0.07)	7.79 (0.06)
	0.18	9.69	1.13	0.92
Buzzati (D)	5.10 (0.03)	17.75 (0.50)	2.67 (0.06)	6.63 (0.11)
	0.20	3.81	0.43	0.86
Buzzati (B)	4.82 (0.02)	15.45 (0.62)	2.41 (0.06)	6.37 (0.15)
	0.15	4.91	0.45	1.16
Calvino (B)	4.63 (0.02)	19.87 (1.02)	2.91 (0.10)	6.73 (0.14)
	0.19	8.61	0.85	1.18

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

[8] The 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).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2018 doi:10.20944/preprints201811.0505.v1

Calvino (M)	4.74 (0.02)	17.60 (0.64)	2.67 (0.08)	6.59 (0.14)
	0.14	3.74	0.49	0.79
Cassola	4.48 (0.02)	11.93 (0.46)	2.11 (0.05)	5.64 (0.10)
	0.18	3.80	0.39	0.86
Collodi	4.60 (0.02)	16.92 (0.60)	2.72 (0.08)	6.19 (0.08)
	0.14	3.83	0.48	0.51
Da Ponte	4.71 (0.01)	26.15 (0.99)	3.78 (0.13)	6.91 (0.08)
	0.17	11.55	1.57	0.93
Deledda	4.51 (0.02)	15.08 (0.64)	2.48 (0.06)	6.06 (0.16)
	0.139	5.03	0.47	1.24
D'Azeglio	4.64 (0.02)	29.77 (1.21)	4.03 (0.13)	7.36 (0.11)
	0.162	11.52	1.25	1.01
De Amicis	4.55 (0.02)	19.43 (0.74)	3.41 (0.11)	5.61 (0.06)
	0.14	6.76	0.97	0.58
De Marchi	4.70 (0.01)	18.95 (0.40)	2.68 (0.05)	7.06 (0.06)
	0.43	4.01	0.46	0.64
D'Annunzio	4.91 (0.16)	17.99 (0.79)	2.79 (0.08)	6.38 (0.16)
	0.112	5.57	0.57	1.15
Eco	4.81 (0.01)	21.08 (0.65)	2.81 (0.07)	7.46 (0.10)
	0.15	8.50	0.85	1.32
Fogazzaro (S)	4.79 (0.01)	14.84 (0.44)	2.34 (0.05)	6.33 (0.10)
	0.10	4.32	0.53	0.94
Fogazzaro (P)	4.79 (0.02)	16.08 (0.42)	2.64 (0.06)	6.10 (0.10)
	0.17	4.43	0.65	1.01
Gadda	4.76 (0.02)	18.43 (1.06)	3.68 (0.17)	4.98 (0.09)
	0.24	10.60	1.70	0.86
Grossi	4.59 (0.01)	28.07 (1.26)	4.23 (0.13)	6.56 (0.13)
	0.14	14.85	1.55	1.55
Leopardi	4.70 (0.03)	31.78 (2.05)	4.54 (0.25)	6.90 (0.23)
	2.62	16.55	2.02	1.20
Levi	4.73 (0.01)	22.94 (0.61)	4.02 (0.11)	5.70 (0.03)
	0.11	5.53	0.96	0.31
Machiavelli	4.71 (0.02)	40.17 (1.02)	6.23 (0.15)	6.45 (0.08)
	0.11	5.39	0.81	0.41
Manzoni (PS)	4.60 (0.02)	24.83 (1.18)	4.63 (0.18)	5.30 (0.06)
	0.32	17.79	2.70	0.95
Manzoni (FL)	4.75 (0.02)	30.98 (1.28)	4.30 (0.14)	7.17 (0.14)
	0.28	19.03	2.10	2.05
Moravia (I)	4.81 (0.02)	36.00 (1.88)	5.34 (0.27)	6.74 (0.09)
	0.15	18.61	2.64	0.93
Moravia (C)	4.56 (0.02)	29.93 (0.87)	4.12 (0.11)	7.28 (0.19)
	0.27	9.73	1.20	2.13
Pavese (B)	4.54 (0.02)	12.37 (0.48)	2.06 (0.05)	5.97 (0.11)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2018 doi:10.20944/preprints201811.0505.v1

	0.09	2.43	0.25	0.56
Pavese (F)	4.47 (0.01)	17.83 (0.67)	2.60 (0.08)	6.83 (0.13)
	0.93	4.42	0.53	0.83
Pellico	4.80 (0.01)	17.27 (0.33)	2.69 (0.06)	6.50 (0.08)
	0.087	2.40	0.41	0.61
Pirandello	4.63 (0.02)	14.57 (0.62)	2.93 (0.08)	4.94 (0.10)
	0.169	5.35	0.69	0.86
Sacchetti	4.37 (0.01)	22.43 (0.58)	3.83 (0.06)	5.82 (0.08)
	0.16	7.66	0.82	1.01
Salernitano	4.40 (0.02)	19.20 (1.18)	3.68 (0.17)	5.14 (0.08)
	0.12	6.96	1.03	0.46
Salgari (C)	4.99 (0.01)	15.09 (0.37)	2.36 (0.04)	6.36 (0.07)
	0.121	3.63	0.38	0.70
Salgari (M)	5.01 (0.01)	15.24 (0.40)	2.44 (0.05)	6.246 (0.08)
	0.11	3.85	0.45	0.74
Svevo	4.86 (0.001)	16.04 (0.58)	2.07 (0.07)	7.75 (0.13)
	0.08	4.76	0.59	1.06
Tomasi di Lampedusa	4.99 (0.02)	26.42 (1.43)	3.33 (0.14)	7.90 (0.14)
	0.17	12.30	1.22	1.23
Verga	4.46 (0.05)	20.45 (0.78)	3.00 (0.10)	6.82 (0.06)
	0.15	7.35	0.97	0.56
Global values	**4.67 (0.006)**	**24.34 (0.29)**	**3.67 (0.04)**	**6.56 (0.03)**

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 (6.3332)	0.993 (2.29)
Calvino (B)	0.997 (4.62)	0.805 (17.21)	0.956 (6.59)	0.9177 (2.6426)
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)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2018 doi:10.20944/preprints201811.0505.v1

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)6	0.809 (3.86)
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)**

Finally, Figure 11 shows the scatter plot between G, G_C, G_F and I_p. We can notice that G_F (and also 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2018 doi:10.20944/preprints201811.0505.v1

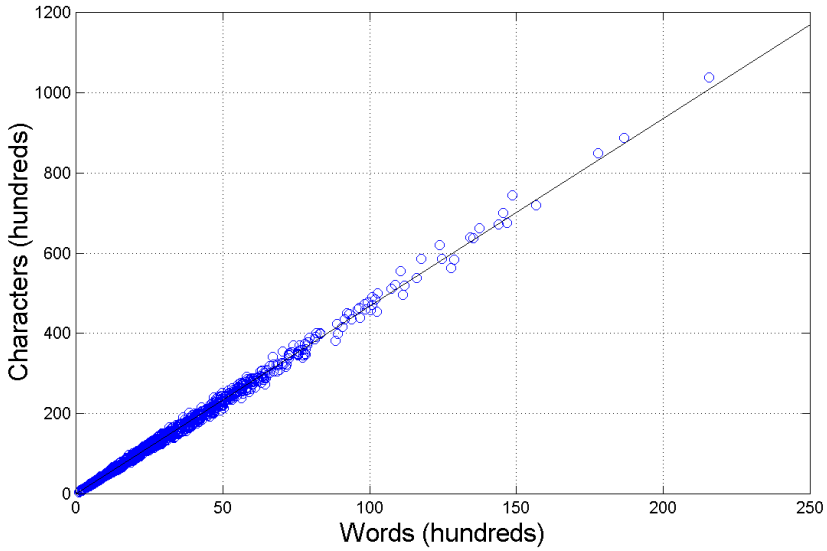

Figure 7: Scatter plot between the number of characters and the number of words (1260 text blocks). Also shown the regression line (see Table 3).

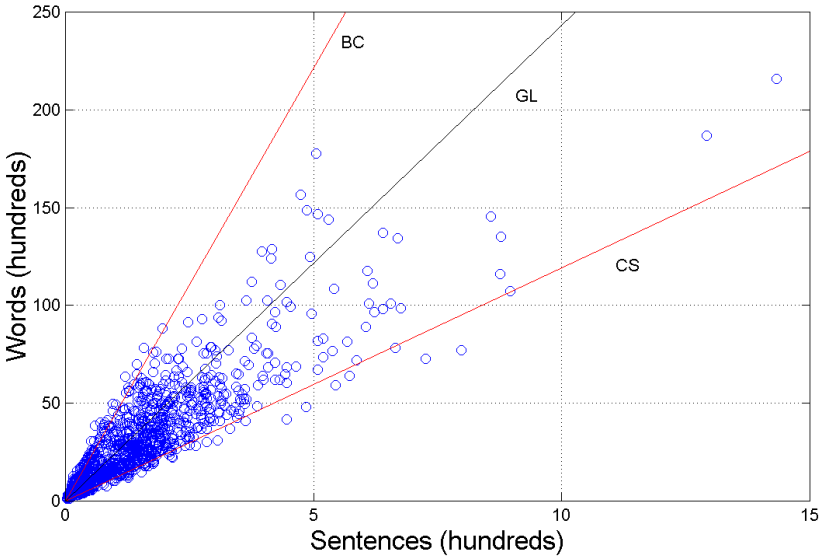

Figure 8: Scatter plot between the number of words and the number of sentences (1260 text blocks). *BC* refers to Boccaccio, *CS* refers to Cassola, *GL* refers to the global values. The two authors represent approximate bounds to the angular region.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2018 doi:10.20944/preprints201811.0505.v1

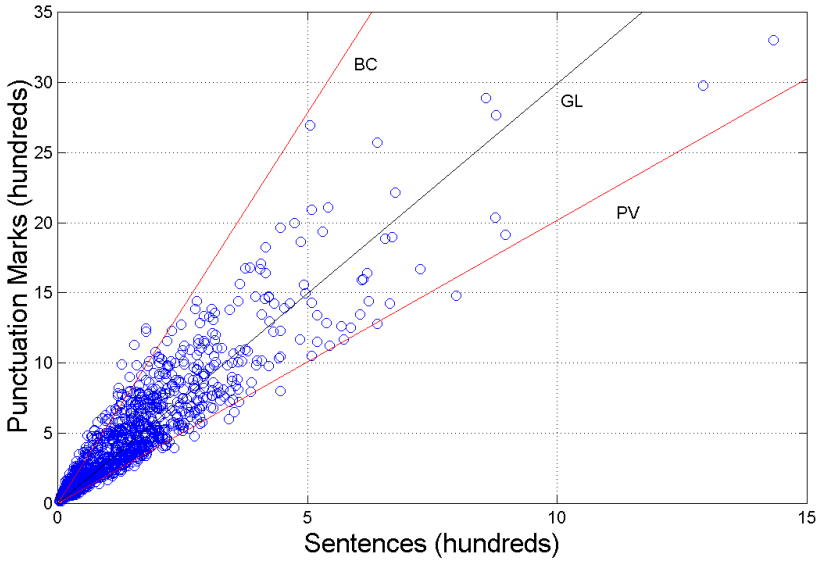

Figure 9: Scatter plot between the number of punctuation marks and the number of sentences (1260 text blocks). *BC* refers to Boccaccio, *PV* refers to Pavese (*La bella estate*), *GL* refers to the global values. The two authors represent approximate bounds to the angular region. The ratio between the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

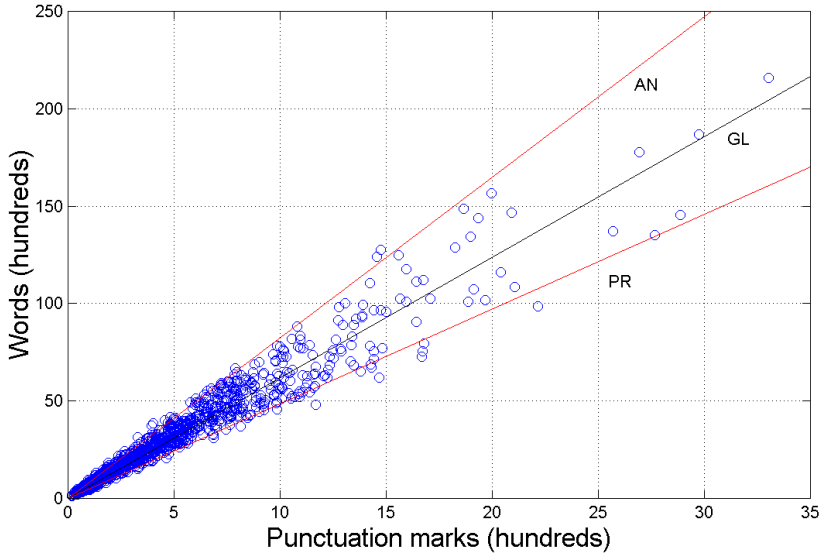

Figure 10: Scatter plot between the number of words and the number of punctuation marks (1260 text blocks). *AN* refers to Anonymous, *PR* refers to Pirandello, *GL* refers to the global values. The two authors represent approximate bounds to the angular region. The ratio between the ordinate and the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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2018 doi:10.20944/preprints201811.0505.v1

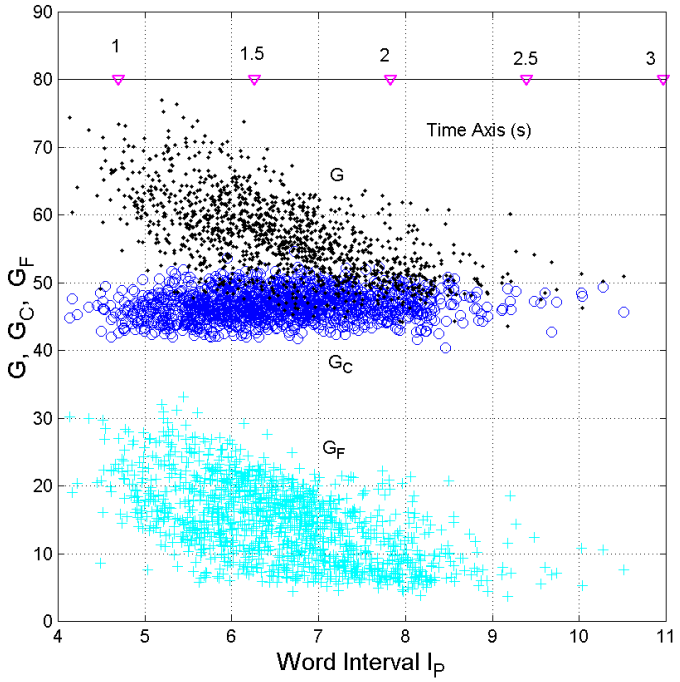

Figure 11: Scatter plot between G (black dots), G_C (blue circles), G_F (cyan crosses) and the word interval I_p. The top time axis refers to the time interval T_P (Section 5).

4. Comparing different literary texts: distances

The large amount 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 traditional approach is to represent documents with $n-$grams using vector representation of particular text features. 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 semantics and syntax are ingnored. A more complex approach represents textual data in more detail (Gómez−Adorno et al., 2016). These new techniques, implemented with complex software, are useful when, together with other tasks, automatic authorship attribution and verification are required.

In the case of the literay texts considered in this paper, we know who the author is and, in my opinion, it is more interesting to compare the statistical characteristcs of different authors or different texts of the same author, by using the data reported in Tables 1,2, 3, instead of using the more complex methods reviewed by (Stamatatos, 2009). For this purpose, the parameters that are most significant are the four random variables defined before: C_P, P_F, M_F and I_P, because they represent fundamental indices and are mostly uncorrelated, except the couple (M_F, 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 components[9] x and y: $\overrightarrow{R_1} = (C_P, P_F)$, $\overrightarrow{R_2} = (M_F, P_F)$, $\overrightarrow{R_3} = (I_P, P_F)$, $\overrightarrow{R_4} = (C_P, M_F)$, $\overrightarrow{R_5} = (I_P, M_F)$, $\overrightarrow{R_6} = (I_P, C_P)$.

[9] The 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2018 doi:10.20944/preprints201811.0505.v1

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

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:

$$S(D_1, D_2) = \frac{1}{6}\sum_{k=1}^{6} \cos(\overrightarrow{R_{D_{1k,}}}, \overrightarrow{R_{D_{2k,}}}) \qquad (9)$$

where $\cos(\overrightarrow{R_{D_{1k,}}}, \overrightarrow{R_{D_{2k,}}})$ is the cosine of the angle formed by the two vectors $\overrightarrow{R_{D_{1k,}}}, \overrightarrow{R_{D_{2k,}}}$. If all pairs of vectors were collinear (aligned), then $\cos(\overrightarrow{R_{D_{1k,}}}, \overrightarrow{R_{D_{2k,}}})$=1, the similarity would be maximum, $S = 1$. If all pairs of vectors were orthogonal $\cos(\overrightarrow{R_{D_{1k,}}}, \overrightarrow{R_{D_{2k,}}})$=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 accepted. 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 literay text (but not with itself). The choice of not excluding the other text blocks of the same literay 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2018 doi:10.20944/preprints201811.0505.v1

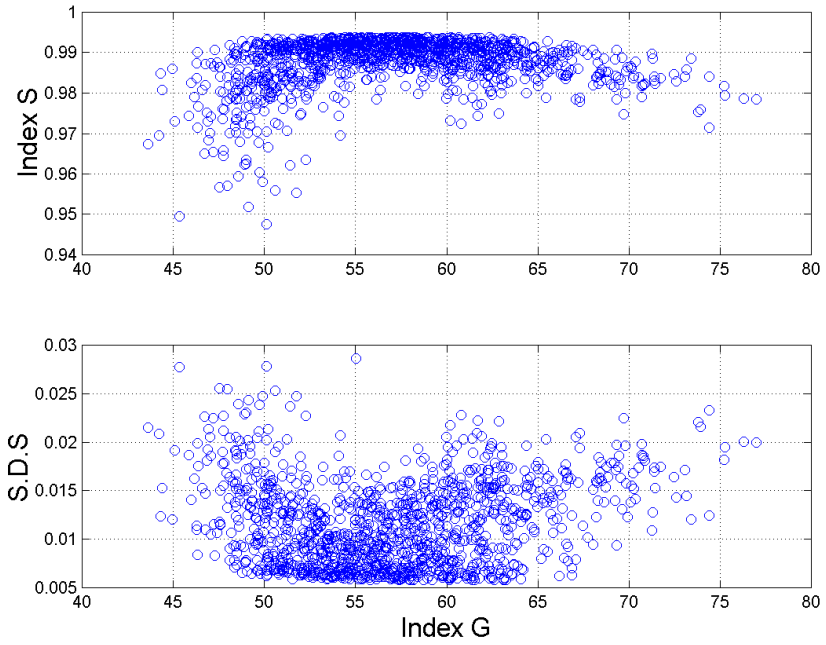

Figure 12: Upper panel: Scatter plot between the average similarity index S of a text block, out of 1260 in total, with regards to all others, and the corresponding readability index G. Lower panel: standard deviation. The total amount of data used to calculate average and standard deviation is given by $1260 \times (1260 - 1) = 1{,}586{,}340$.

Now, a better choice for comparing literary texts is to consider the "distance" of any text block from the origin of x and y axes[10], given by the magnitude of the resulting vector $\vec{R}$:

$$R = \left|\vec{R}\right| = \left|\sum_{k=1}^{6} \overrightarrow{R_k}\right| \qquad (10)$$

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 literay texts (obtainable from Tables 1,2,3) are also shown.

It is very interesting, for example, to compare the vector representing Manzoni's masterpiece[11] *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 experts 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 106.6[12]. Other interesting observations are the coincidence of the two vectors representing the novels by Fogazzaro, and the difference between the novels by Pavese, etc.

---

[10] From vector analysis, the two components of a vector are 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 $R = \sqrt{x^2 + y^2}$.

[11] A compulsory reading in any Italian High School.

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2018 doi:10.20944/preprints201811.0505.v1

With this tool, the experts 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.

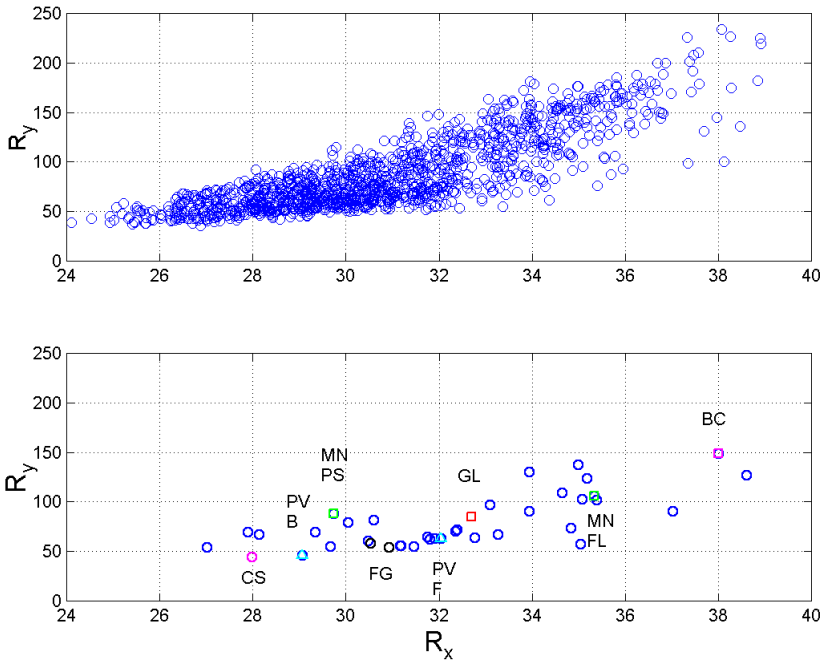

Figure 13: Scatter plot between the two components of the distance R for all 1260 text block (upper panel), and that calculated from the average values shown in Tables 1,2,3 (lower panel). *CS*=Cassola, *PV B*=Pavese *La bella estate*, *PV F*= Pavese *La luna e i falò*, MN PS=Manzoni *I promessi sposi*, MN FL=Manzoni Fermo e Lucia, FG= Fogazzaro *Il santo* and *Piccolo mondo antico*, BC=Boccaccio, GL=global values ("barycentre").

The efficay of $\vec{R}$ can be appreciated in Figure 14, which shows the scatter plot between R and G, and between its angle $\varphi = tan^{-1}(\frac{y}{x})$ and G. The black lines describes very well the relationships between them, given by:

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

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

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, that between measured and estimated values of φ with (11b) is 0.867[13]. In conclusion, the magnitude (distance) R and the angle φ of the vector $\vec{R}$ are very well correlated with the readability index G.

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2018 doi:10.20944/preprints201811.0505.v1

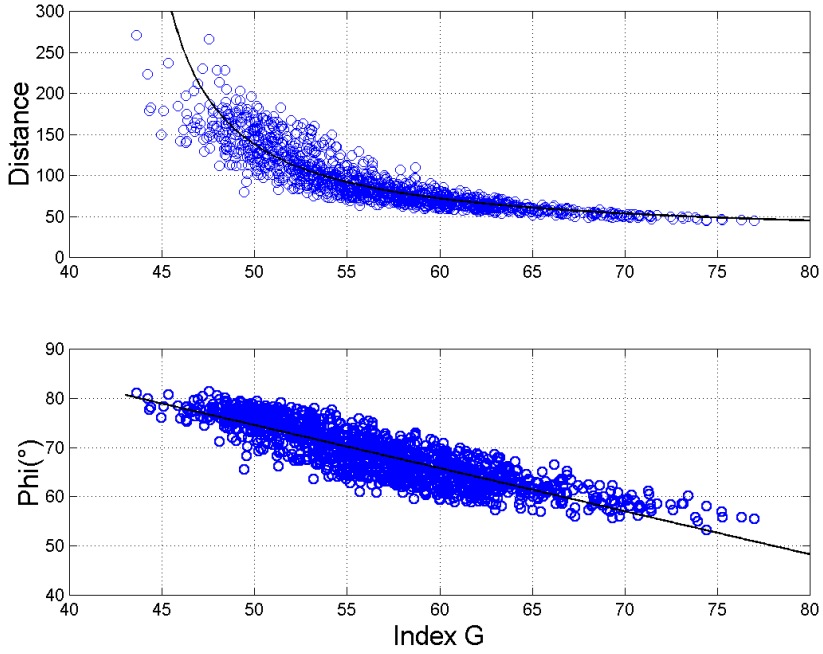

Figure 14: Scatter plot between R and G (upper panel) and between $\varphi = tan^{-1}(\frac{y}{x})$ and G (lower panel).

5. Word interval and Miller's 7 ∓ 2 law

The range in which the word interval I_p varies, 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 (Cowan, 2000), (Bachelder, 2000), (Chen and Cowan, 2005), (Mathy and Feldman, 2012), (Gignac, 2015). For words, i.e. for data that can be restricted (i.e., "compressed") by chunking, it seems that the average value is not 7 but around 5 to 6 (Miller, 1955), 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 (Gignac, 2015), 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, practically 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 0.975 and 0.025) fall in the range from 4.6 to 8.6, which concides, 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[14]:

[14] 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 (Bury, 1975) 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; $\mu_{I_P} = log\left[\left(m_{I_P} - 1\right) - \frac{\sigma_{I_P}^2}{2}\right]$=1.698.

The mode (the most likely value) is given by $\mu_o = exp\left(\mu_{I_P} - \sigma_{I_P}^2\right) + 1 = 6.297$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2018 doi:10.20944/preprints201811.0505.v1

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

with confidence level in excess of 99.99% (chi-square test) (Papoulis, 1990). 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).

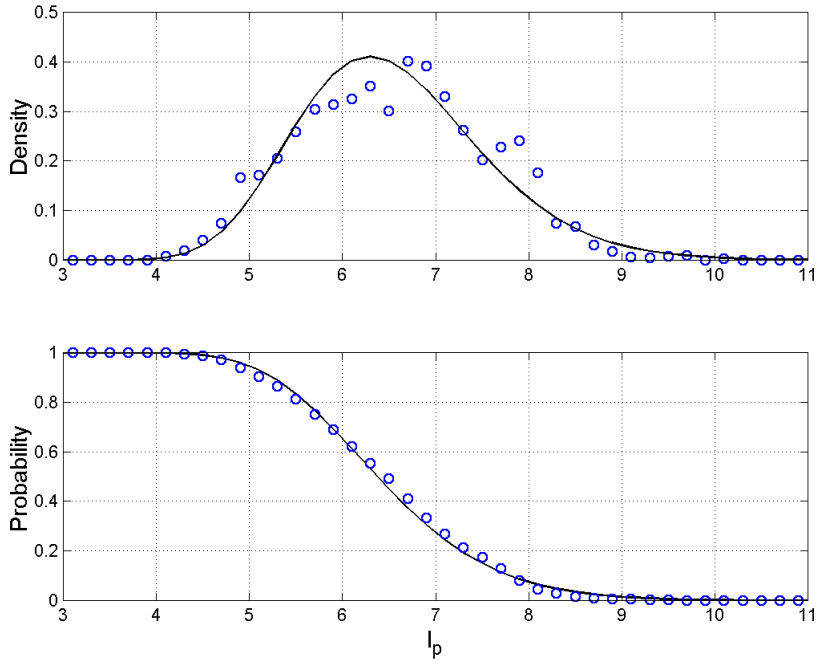

Figure 15: Probability density function (upper panel, blue circles) and the complementary probability distribution (lower panel, blue circles) of I_P for 1260 text bloks. The lower panel shows the probability that the value reported in abscissa (x axis) is exceeded. The black continuous lines are the theoretical density and distribution of a three−parameter log−normal model (Bury, 1975).

Theses 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 (Jarvella, 1971). 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, with the result that the text is less readable, a result quantitatively expressed by the empirical equation (1) for Italian.

In conclusion, the range of the word interval is similar to Miller's law range. The values found for each author, in our opinion, sets 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 will be found reliable after more studies concerning short-term memory and brain, the link between G_F, and hence G through equation (6), would appear justified and natural.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2018 doi:10.20944/preprints201811.0505.v1

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 (Trauzettel−Klosinski and Dietz, 2012). In this case, the average time interval corresponding to the word interval, expressed in seconds, is given by:

$$I_T = 60 \times \frac{I_P}{188} \qquad (12)$$

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 (Baddeley et al., 1975), (Mandler and Shebo, 1982), (Muter, 2000) (Grondin, 2000), (Pothos and Joula, 2000), (Chekaf et al., 2016).

In my opinion, these results, relatin I_T and I_P to fundamental and accessible characteristics of short−term memory, are very interesting and should be furtherly pursued by experts, not by this author. 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.

7. 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 readabilty 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_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. 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.

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, in my opinion, that between the short−term memory capacity, described by Miller's "7 ∓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 (or 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 literay 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 flavour 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2018 doi:10.20944/preprints201811.0505.v1

similar to those reported in this paper, therefore evidencing some universal and long−lasting characteristics of western languages and their readers. These results will be reported elesewhere.

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 (Conway et al., 2002), this relationship seems to be very useful. However, its relationship with Miller's law should be further investigated because, in my opinion, the word interval is another parameter that can be used to design a text, together with readability formulae, to better match expected reader's characteristics.

Technical and scientific writings (papers, essays etc.) ask more to their readers. A preliminary investigation done on short scientific texts published in the Italian popular science magazines *Le Scienze* and *Sapere* (today is rare to find original scientific papers written in Italian), in a popular scientific book and newspaper editorials give 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 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, deals with abstract meaning with articulation syntactically elaborated, and leading to long sentences comprising series of subordinate clauses. Of course, the reader of these texts expects to find technical and abstarct terms of his field, or specialty, and would not understand the text if these elements were absent.

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[15]	188 536	36 122	1505	49.31	52.19	12.50	5.22	24.23	2.86	8.51
				(0.53)	(0.36)	(0.25)	(0.04)	(0.52)	(0.06)	(0.15)
Popular scientific articles[16]	188 537	33 797	1241	44.23	55.79	11.02	5.58	27.97	2.95	9.57
				(0.63)	(0.37)	(0.37)	(0.04)	(0.90)	(0.11)	(0.24)
Newspapers editorials[17]	60281	11919	479	50.48	50.58	12.06	5.06	25.28	2.99	8.67
				(0.59)	(0.47)	(0.45)	(0.05)	(0.88)	(0.16)	(0.44)

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

[16] *Le Scienze*, *Scienze e ricerche*, 2017 issues.

[17] *Il Corriere della Sera*, *La Repubblica*, *Il Sole 24 ore*, 2018.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2018 doi:10.20944/preprints201811.0505.v1

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 readabilty index
G_s	scaled readabilty 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

Conflicts of Interest: The author declares no conflict of interest.

References

(Anderson, 1991) Anderson, P.V., *Technical Writing: a Reader−Centered Approach* (1991), Harcourt Brace Jovanovich, Fort Worth, Texas, 2nd edition.

(Atvars, 2016). Atvars, A. (2016), Eye Movement Analyses for Obtaining Readability Formula for Latvian Texts for Primary School, *Procedia Computer Science* , 104, 477− 484.

(Bachelder, 2000), Bachelder, B-L., The magical number 4=7: Span theory on capacity limitations, *Behavioral and Brain Sciences*, 24, 116−117.

(Baddeley et al., 1975), Baddeley, A.D., Thomson, N., Buchanan, M. (1975), Word Length and the Structure of Short−Term Memory, Journal of Verbal Learning and Verbal Behavior, 14, 575−589.

(Bailin and Graftstein, 2001), Bailin, A., Graftstein, A. (2001), The linguistic assumptions underlying readability formulae: a critique, *Language & Communication*, 21, 285−301.

(Barrouillest and Camos, 2012), Barrouillest, P., Camos, V. (2012), As Time Goes By: Temporal Constraints in Working Memory, Currenti Directions in Psychological Science, 21(6), 413–419.

(Benjamin, 2012), Benjamin, R.G. (2012), Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty, *Educ Psycological Review*, 24, 63–88.

(Burnett, 2004) Burnett R.E., *Technical Communication* (2004), Wadsworth Publishing Company, Belmont, California, 3rd edition.

(Bury, 1975), Bury, K.V. (1975), *Statistical Models in Applied Science*, John Wiley.

(Chekaf et al., 2016) Chekaf, M., Cowan, N., Mathy, F. (2016), Chunk formation in immediate memory and how it relates to data compression, *Cognition*, 155, 96–107.

(Chen and Cowan, 2005), Chen, Z., Cowan, N. (2005), Chunk Limits and Length Limits in Immediate Recall: A Reconciliation, *J. Exp. Psychol. Mem. Cogn*., 3, 1235–1249.

(Collins–Thompson, 2014), 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*, 165:2, 97–135.

(Conway et al., 2002) Conway, A.R.A., Cowan, N., Michael F. Bunting, M.F., Therriaulta, D.J. (2002), Minkoff, S.R.B., A latent variable analysis of working memory capacity, short–term memory capacity, processing speed, and general fluid intelligence, *Intelligence*, 30, 163–183.

(Cowan, 2000), Cowan, N. (2000), The magical number 4 in short–term memory: A reconsideration of mental storage capacity, *Behavioral and Brain Sciences*, 24, 87–114.

(De Mauro, 1980), De Mauro, T. (1980), *Guida all'uso delle parole*, Editori Riuniti, Roma.

(DuBay, 2004) DuBay, W.H. (2004), *The Principles of Readability*, Impact Information, Costa Mesa, California.

(DuBay, 2006) DuBay (Editor), W.H. (2006), *The Classic Readability Studies*, Impact Information, Costa Mesa, California.

(Gignac, 2015) 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.

(Grondin, 2000) Grondin, S. (2000), A temporal account of the limited processing capacity, *Behavioral and Brain Sciences*, 24, 122–123.

(Grzybeck, 2007) Grzybeck, P. (2007), History and methodology of word length studies, *Contributions to the Science of Text and Language*, Dordrecht: Spinger, 15–90.

(Jarvella, 1971), Jarvella, R.J. (1971), Syntactic Processing of Connected Speech, *Journal of Verbal Learning and Verbal Behavior*, 10(4), 409–416.

(Jones and Macken, 2015) Jones, G, Macken, B., (2015), Questioning short–term memory and its measurements: Why digit span measures long–term associative learning, *Cognition*, 144, 1–13.

(Lucisano and Piemontese, 1988) Lucisano, P., Piemontese, M.E. (1988), GULPEASE: una formula per la predizione della difficoltà dei testi in lingua italiana, *Scuola e città*, 110–124.

(Mandler and Shebo, 1982) Mandler G, Shebo, B.J. (1982), Subitizing: An Analysis of Its Componenti Processes, Journal of Experimental Psychology: General, III(1), 1–22.

(Maraschio, 1993) Maraschio, N. (1993), Grafia e ortografia: evoluzione e codificazione, *Storia della lingua italiana*: *I luoghi della codificazione*, Einaudi,Torino, I, 139–227.

(Martin and Gottron, 2012), Martin, L., Gottron, T. (2012), Readability and the Web, *Future Internet*, 4, 238–252.

(Mathy and Feldman, 2012) Mathy, F., Feldman, J. (2012), What's magic about magic numbers? Chunking and data compression in short−term memory, *Cognition*, 122, 346−362.

(Matricciani, 2007) Matricciani, E. *La scrittura tecnico−scientifica* (2007), Casa Editrice Ambrosiana, Milano.

(Miller, 1955), Miller, G.A. (1955), The Magical Number Seven, Plus or Minus Two. Some Limits on Our Capacity for Processing Information, *Psychological Review*, 343−352.

(Mortara Garavelli, 2003), Mortara Garavelli, B. (2003), *Prontuario di punteggiatura*, Editori Laterza

(Muter, 2000) Muter, P. (2000), The nature of forgetting from short−term memory, *Behavioral and Brain Sciences*, 24, 134.

(Papoulis, 1990) Papoulis, A. (1990), Probability & Statistics, Prentice Hall.

- (Parkes, 2016) Parkes, M.B. (2016), *Pause and Effect. An Introduction to the History of Punctuation in the West*, Routledge, 343 pages.

(Pothos and Joula, 2000), Pothos, E.M., Joula, P. (2000), Linguistic structure and short−term memory, *Behavioral and Brain Sciences*, 24, 138−139.

(Serianni, 2001), Serianni, L. (2001), Sul punto e virgola nell'italiano contemporaneo, *Studi Linguistici italiani*, XXVII, 2, 248−255.

(Saaty and Ozdemir, 2003) Saaty, T.L., Ozdemir, M.S. (2003), Why the Magic Number Seven Plus or Minus Two, *Mathematical and Computer Modelling*, 38, 233−244.

(Stamatatos, 2009), Stamatatos, E. (2009), A survey of Modern Authorship Attribution methods, Journal of the American Society for Information Science and Technology, 60(3), 538−556.

(Trauzettel−Klosinski and Dietz, 2012), Trauzettel−Klosinski, S., K. Dietz, K. (2012), Standardized Assessment of Reading Performance:The New International Reading Speed Texts IreST, *IOVS*, 5452−5461.

(Vajjala et al., 2016), Vajjala, S., Meurers, D., Eitel, A., 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, Japan, December 11−17, 38−48.

(Zamanian and Heydari, 2012), Zamanian, M., Heydari, P. (2012), Readability of Texts: State of the Art, *Theory and Practice in Language Studies*, 2, 43−53