The Theory of Linguistic Channels in Alphabetical Texts

The Theory of Linguistic Channels in Alphabetical Texts

By

Emilio Matricciani

Cambridge
Scholars
Publishing

The Theory of Linguistic Channels in Alphabetical Texts

By Emilio Matricciani

This book first published 2024

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-5275-5528-3
ISBN (13): 978-1-5275-5528-0

TABLE OF CONTENTS

FOREWORD

The book is a collection of papers, published in the last five years, on an original mathematical/statistical theory concerning the deep–language mathematical surface structure of alphabetical texts. The theory does not follow the actual paradigm of linguistic studies, which neither consider Shannon's communication theory, nor the fundamental connection that some linguistic parameters have with reading skill and short–term memory capacity of readers.

The development of my research appears in the temporal sequence of the papers, therefore, the reader will find, inevitably, some repetitions – useful, at the time of publication, to make paper reading easier – and some slightly different mathematical symbols.

In collecting these papers, my aim is to propose to young researches and students – in the fields of cognitive psychology, theory of communication, information theory, phonics and linguistics, history of modern and ancient literatures, stylometry – a possible theoretical framework which could allow the brightest of them to further research the fundamental mathematical structure of human language. This research might lead, in the end, to devise a mathematical theory that includes meaning, the great absent element – since Shannon's times – in our mathematical theories of human communication.

In the following, I sketch a brief summary of each chapter. The reader can read them in the indicated sequence – recommended – or according to his/her interests and curiosity. The material can be used for a series of seminars or for a trimester course in graduate classes.

Chapter 1. *Deep Language Statistics of Italian throughout seven centuries of Literature and connection with miller's* 7 ± 2 *Law and short–term memory.*

In this chapter I lay down the first elements of the mathematical/statistical theory. 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 shows that of the two terms that determine the readability index – the *semantic index*, proportional to the number of characters per word, and the *syntactic index*, proportional to the reciprocal of the number of words per sentence, the latter is dominant. Each author can modulate the length of sentences more freely than he/she can do with the length of words, and in different ways from author to author. A universal readability index is proposed in Chapter 6.

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 in a language. The most important relationship 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* 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 it is further investigated in the following chapters.

Chapter 2. *A Statistical Theory of Language Translation Based on Communication Theory.*

In this chapter, I propose the first statistical theory of language translation based on communication theory. The theory is based on New Testament translations from Greek to Latin and to other 35 modern languages. In a text translated into another language, all linguistic variables do numerically change. To study the chaotic data that emerge, I model any translation as a complex communication channel affected by "noise", studied according to Communication Theory applied for the first time. Notice that this theory can be applied to any other field of research if regression lines are considered, therefore it is of general interest.

The input language is the "signal", the output language is a "replica" of the input language, but largely perturbed by noise, indispensable, however, for conveying the meaning of the input language to its readers. I define a noise–to–signal power ratio and show that linguistics channels – suitably defined – are differently affected by translation noise. I propose a global readability formula for alphabetical languages – not available for most of them – and conclude with a general theory of language translation which shows that direct and reverse linguistic channels are not symmetric. The general theory can also be applied to linguistic channels of texts belonging to the same language both to study how texts of the same author may have changed over time, or to compare texts of different authors, which falls in the domain of stylometry.

Chapter 3. *Linguistic Mathematical Relationships Saved or Lost in Translating Texts: Extension of the Statistical Theory of Translation and Its Application to the New Testament.*

In this chapter I extend the general theory of translation to texts written in the same language. The main result shows how much the mutual mathematical relationships of texts in a language have been saved or lost in translating them into another language. To make objective comparisons, I define a “likeness index” I_L – based on probability and communication theory of noisy binary digital channels – and show that it can reveal similarities and differences of texts. I apply the theory to the New Testament translations in Latin and in 35 modern languages. To avoid the inaccuracy, due to the small sample size from which the input data (regression lines) are calculated, I have adopted a “renormalization” based on Monte Carlo simulations whose results can be considered as “experimental”. In general, in many languages/translations the original linguistic relationships are lost and texts are mathematically distorted.

Chapter 4. *Multiple Communication Channels in Literary Texts.*

In this chapter, I use the theory to compare how a literary character speaks to different audiences by diversifying two important linguistic communication channels: the “sentences channel” and the “interpunctions channel”. The theory can “measure” how the author shapes a character speaking to different audiences, by modulating deep–language surface parameters. To show its power, I have applied the theory to the literary corpus of Maria Valtorta, an Italian mystic of the XX–century.

The likeness index I_L , ranging from 0 to 1, allows to “measure” how two linguistic channels are similar, therefore implying that a character speaks to different audiences in the same way.

Chapter 5. *Capacity of Linguistic Communication Channels in Literary Texts. Application to Charles Dickens’ Novels.*

In the first part of the chapter I recall the general theory of linguistic channels – based on regression lines between deep–language surface parameters – and study their capacity and interdependence. In the second part I apply the theory to novels written by Charles Dickens and by other authors of the English Literature, including the Gospels in the King James version.

Dickens’ novels show a striking and unexpected mathematical/statistical similarity with the synoptic Gospels.

Chapter 6. *Readability Indices Do Not Say It All on a Text Readability.*

In this chapter, I propose a universal readability index, G_U, applicable to any alphabetical language and related to the cognitive psychology, theory of communication, phonics and linguistics. This index considers also readers' short−term memory processing capacity, here modeled by the word interval I_P, namely the number of words between two contiguous interpunctions.

Any current readability formula does not consider I_p, but scatterplots of I_p versus a readability index show that texts with the same readability index can have very different I_p. It is unlikely that I_P has no impact on reading difficulty. The examples shown are taken from Italian and English Literatures, and from the translations of New Testament in Latin and in 35 modern languages.

Chapter 7. *Short–Term Memory Capacity Across Time and Language Estimated from Ancient and Modern Literary Texts. Study–Case: New Testament Translations.*

In this chapter, I study the short−term memory capacity of ancient readers of the original New Testament written in Greek, and that of readers of its translations in Latin and in 35 modern languages. To model the short−term memory, I consider the number of words between any two contiguous interpunctions I_p.

Since I_P can be calculated for any alphabetical text, we can perform experiments − otherwise impossible − with ancient readers by studying the literary works they used to read.

Chapter 8. *Readability Across Time and Languages: the case of Matthew's Gospel Translations.*

In this chapter, I study how the readability of a text can change in translation by considering Matthew's Gospel, written in Greek, translated in Latin and in 35 modern languages. I have found that the deep−language surface parameters of each translation are so diverse from language to language, and even within a given language for which there are many versions of Matthew − as in English and in Spanish – that the resulting texts mathematically seem to be diverse texts.

The several tens of versions of Matthew's Gospel studied appear to address very diverse audiences. If a reader could understand well all of them, he/she would get the impression of reading texts written by diverse authors, although all of them tell the same story.

Chapter 9. *Linguistic Communication Channels Reveal Connections Between Texts: The New Testament and Greek Literature*.

In this chapter, I study two fundamental linguistic channels – namely the sentence channel and the interpunctions channel – and show that they can reveal deeper connections between texts. As study–case, I have considered the Greek New Testament, with the purpose of determining mathematical connections between its texts and possible differences in writing style (mathematically defined) of writers, and in reading skill required to their readers. To set the New Testament texts in the Greek Classical Literature, I have also studied and compared texts written by Aesop, Polybius, Flavius Josephus and Plutarch.

I have found large similarity (measured by the likeness index of Chapter 3) in the couples of texts *Luke–Matthew*, *Luke–Mark*, *John–Matthew*, *John–Mark*, *Luke–Acts*, findings that largely confirm what scholars have found about these texts, therefore giving credibility to the theory. The gospel according to *John* is very similar to Aesop' *Fables*. *John* might have been inspired by the long tradition of short stories telling a truth, such as *Fables.* Surprisingly, *Hebrews* and *Apocalypse* are each other "photocopy" in the two linguistic channels, and not linked to all other texts.

Chapter 10. *Is Short–Term Memory Made of Two Processing Units? Clues from Italian and English Literatures Down Several Centuries*.

In this chapter, I conjecture that the short−term memory processing to convey the meaning of a sentence is made by two uncorrelated processing units in series. The clues for conjecturing this model emerge from studying many novels of the Italian and English Literatures, already considered in Chapter 6. I show that there are no significant mathematical/statistical differences between the two literary corpora. This finding means that the mathematical structure of alphabetical languages is very deeply rooted in human mind, independently of the language used.

The first processing unit is linked to the number of words between two contiguous interpunctions, variable I_p, approximately ranging in Miller's 7 ± 2 range; the second unit is linked to the number of I_p's contained in a sentence, variable M_F, ranging approximately from 1 to 6.

Finally, I wish to thank the many scholars who, with great care and love, maintain digital texts available to readers and scholars of different academic disciplines, such as Perseus Digital Library and Project Gutenberg.

Politecnico di Milano, Milan, Italy
September 2023.

CHAPTER 1

DEEP LANGUAGE STATISTICS OF ITALIAN THROUGHOUT SEVEN CENTURIES OF LITERATURE AND CONNECTION WITH MILLER'S 7 ± 2 LAW AND SHORT−TERM MEMORY[1]

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* is spread in the same of Miller's law, and the *time*

[1] 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. doi: 10.4236/ojs.2019.93026.

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.

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 [4−6], 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 [3,7−9], 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 [10−13]. 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 consider 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 1980’s [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 – especially − 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 (interpunctions) 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 chapter, 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 [17]. 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 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.

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 – languages 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 chapter is to research, with regard to the high Italian language, the following topics:

a) The impact of character and sentence 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 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.

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:

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

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 Eq. (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). Let us define the terms:

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

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

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

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

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, 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 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 writers[2], 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[3].

3. Relationships among G_C , G_F and G revisiting the GULPEASE readability formula of Italian

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 P_F, 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 or

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

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

notes. 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.

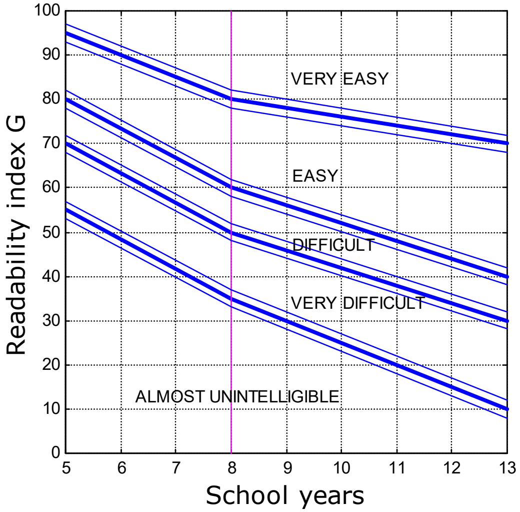

Figure 1. Readability index G of Italian, as a function of the number of school years attended (in Italy high school lasts 5 years, children attend it up to 19 years old). The blue thinner lines indicate the error bounds found in using Eq. (6). Elementary school lasts 5 years, Scuola Media Inferiore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

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

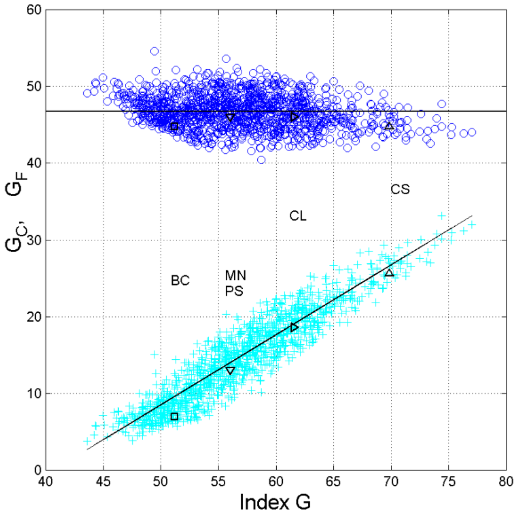

Figure 2a: G_C (blue circles) and G_F (cyan crosses) versus G, for all 1260 text blocks. BC=Boccaccio, MN PS=Manzoni (*I promessi sposi*), CL=Collodi, CS=Cassola. The horizontal line is the average value of G_C, equation (3), the diagonal line is the average value of G_F, equation (4).

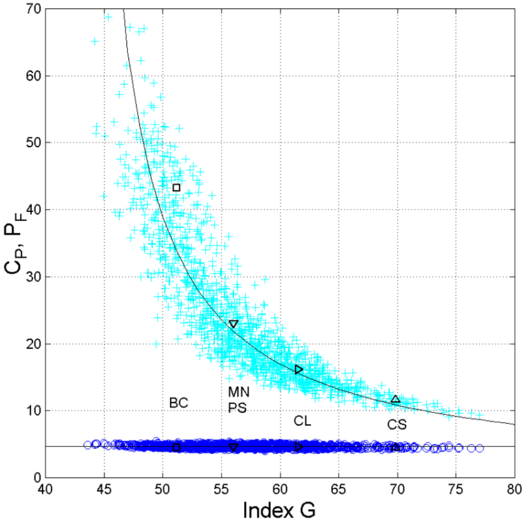

Figure 2b: Scatter plots of $C_P = c/p$ vs. G (blu dots) and $P_F = p/f$ vs. G (cyan crosses). BC=Boccaccio, MN PS=Manzoni (*I promessi sposi*), CL=Collodi, CS=Cassola. The black continuous lines are given by Eqs. (3) and (5).

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 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.

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 5.87	47.14 0.14 1.65	11.95 0.43 5.04
Deledda Grazia (*Canne al vento*, XX, Nobel Prize 1926)	276552	61375	4184	64.39 0.92 7.21	45.06 0.18 1.39	20.45 0.79 6.22

[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 $\sigma_r = \sigma \times \sqrt{\frac{n_r}{n}}$=$\sigma_\mu\sqrt{n_r}$.

D’Azeglio Massimo (*Ettore Fieramosca*, XIX)	424259	91464	3182	53.05 0.54 5.15	46.39 0.17 1.62	10.44 0.45 4.33
De Amicis Edmondo (*Cuore*, XIX)	376792	82770	4775	60.78 0.55 5.01	45.52 0.15 1.39	17.31 0.54 4.94
De Marchi Emilio (*Demetrio Panelli*, XIX)	471451	100328	5363	58.05 0.44 4.37	46.99 0.14 1.43	16.04 0.35 3.48
D’Annunzio Gabriele (*Le novelle delle Pescara*, XX)	244055	49688	3027	58.16 0.81 5.74	49.12 0.16 1.12	18.28 0.77 5.44
Eco Umberto (*Il nome della rosa*, XX)	821707	170676	8490	55.78 0.52 6.82	48.14 0.11 1.47	14.92 0.49 6.35
Fogazzaro (*Il santo*, XIX−XX)	467990	97616	6637	61.46 0.72 7.16	47.94 0.10 1.02	20.40 0.65 6.40
Fogazzaro (*Piccolo mondo antico*, XIX−XX)	528659	112706	7069	61.46 0.57 6.06	47.94 0.16 1.65	20.40 0.48 5.12
Gadda (*Quer pasticciaccio brutto*… XX)	474359	99631	5596	58.24 1.08 10.77	47.61 0.24 2.35	16.85 1.00 9.99
Grossi Tommaso (*Marco Visconti*, XIX)	637311	138900	5301	54.57 0.63 7.39	45.88 0.12 1.42	11.45 0.55 6.53
Leopardi Giacomo (*Operette morali*, XIX)	322991	68699	2694	53.76 1.17 8.46	47.01 0.32 2.62	11.77 0.96 6.46
Levi (*Cristo si è fermato a Eboli*)	383893	81092	3611	55.02 0.39 3.54	47.34 0.13 1.13	13.36 0.36 3.27
Machiavelli Niccolò (*Il principe*, XV−XVI)	130274	27680	702	49.54 0.33 1.75	47.06 0.21 1.10	7.61 0.21 1.09
Manzoni Alessandro (*I promessi sposi*, XIX)	1036728	225392	9766	56.00 0.69 10.29	46.00 0.21 3.18	13.00 0.52 7.76
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	59.420.51 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 0.16	46.70 0.06	14.40 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 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 2b, 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 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):

$$G_C = 46.7 \tag{3}$$

and the average value of G_F described by the regression line

$$G_F = 0.912 \times G - 37.1 \tag{4}$$

The correlation coefficient between G_F and G, Eq. (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 2a shows also the average values of selected works listed in Table 1 to locate them in this scatter plot. Figure 2b shows, superposed to the scattered values of P_F, the theoretical relationship between the average value of P_F, as a function of G, given, according to Eqs. (1a) and (3), by:

$$P_F = \frac{300}{G-42.3} \tag{5}$$

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, Eq. (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:

$$G_{estim} = G_{min} + G_F \tag{6}$$

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 ampler 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 ∓6.8% change. On the contrary G_F varies from 6.94 (Boccaccio) to 25.65 (Cassola), with excursions in the range −52% to ∓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, 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].

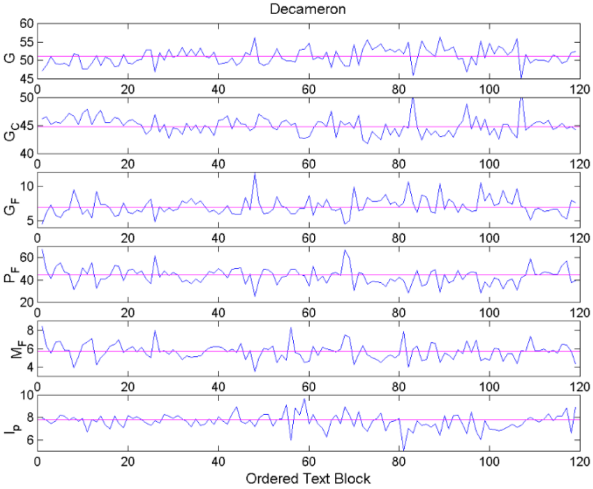

Figure 3: Ordered text−block series of G, G_C, G_F, P_F, M_F and I_p, versus the ordered sequence of text blocks found in Boccaccio's *Decameron*. The text blocks are the novels told, on turn, by each character each day. The horizontal magenta lines give the average values (Tables 1, 2)

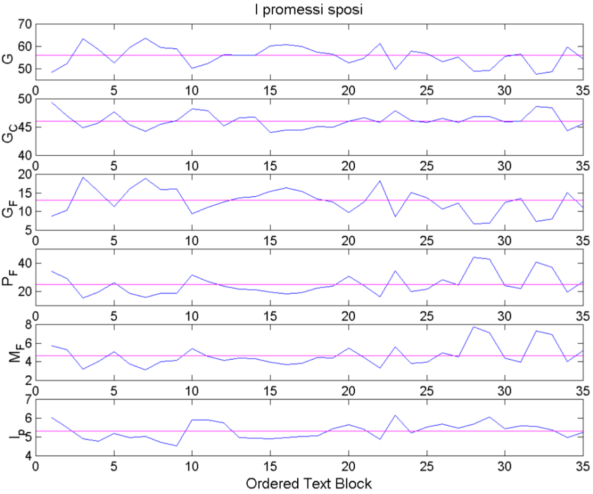

Figure 4: Ordered text–block series series of G, G_C, G_F, P_F, M_F and I_p, versus the ordered sequence of text blocks found in Manzoni's *I promessi sposi*. The text blocks are the chapters. The horizontal magenta lines give the average values (Tables 1, 2).

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 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 Eq. (1a), 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 constant value G_{min} can be set to zero, therefore making:

$$G \cong G_S = G_F \qquad (7)$$

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

with the advantage that the scaled index G_s starts at 0. Now Eq. (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 particular effort. Eq. (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).

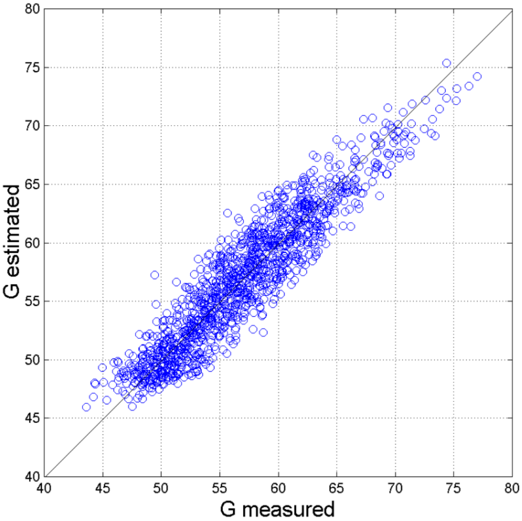

Figure 5a: Scatter plot of G_{estim} equation (6), and the values G calculated with equation (1). Also shown the regression line, in practice the 45° line $G = G_{estim}$.

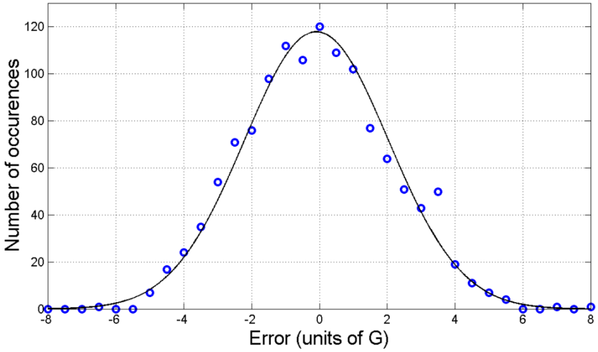

Figure 5b: Histogram of the error $G - G_{estim}$ (blue circles) and theoretical histogram (black line) due to a Gaussian (normal) density function with average value -0.1 and standard deviation 2.14.

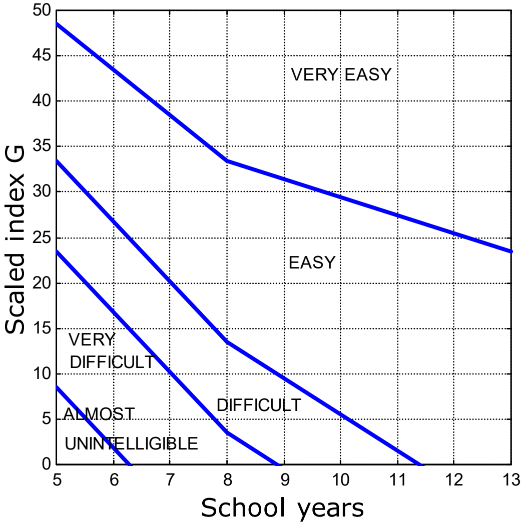

Figure 6: Scaled index G_S as a function of the number of school years attended in the Italia School System.