Preprints.org

Translation Can Distort the Linguistic Parameters of Source Texts Written in Inflected Language. Case Study: Multidimensional Mathematical Analysis of "The Betrothed", a Translation in English of "I Promessi Sposi" by A. Manzoni

Emilio Matricciani [*]

Article

Translation Can Distort the Linguistic Parameters of Source Texts Written in Inflected Language. Case Study: Multidimensional Mathematical Analysis of "The Betrothed", a Translation in English of "I Promessi Sposi" by A. Manzoni

Emilio Matricciani

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

Abstract: We compare, mathematically, the text of a famous Italian novel, *I promessi sposi* written by Alessandro Manzoni (source text), to its most recent English translation, *The Betrothed* by Michael F. Moore (target text). The mathematical theory applied does not measure the efficacy and beauty of texts, but only their mathematical underlying structure and similarity. The translation theory adopted by the translator is the "domestication" of the source text. The translator justified the almost total replacement of semicolons with periods because English is not as economical in its use of subject pronouns as Italian. Question and exclamation marks, and commas too, were largely modified. These modifications produce several consequences. The short-term memory buffers required to the reader are diverse in Italian and in English. The number of sentence patterns allowed in Italian is about two orders of magnitude greater than in English. The geometrical representation of texts and the related probability of error indicates the two texts are practically uncorrelated. All linguistic channels are very noisy, with very poor signal-to-noise ratio, except the channel related to characters and words. Readability indices are also diverse. In conclusion, a blind comparison of the linguistic parameters of the two texts would unlikely indicate they refer to the same novel.

Keywords: alphabetical languages; short–term memory; geometrical representation; linguistic variables; linguistic channels; probabilitty of error; universal readability index

1. Introduction

Translation is the replacement of a text in one language (source text) by an equivalent text in another language (target text). Most studies on translation report results not based on mathematical analysis of texts, as we do with our mathemartical/statistical theory on alphabetical languages developed in a series of papers [1–8]. When a mathematical approach is used [9–13], scholars neither consider Shannon's communication theory [14] nor the fundamental connection linguistic variables have with reader's reading skill and short–term memory, considered instead by our theory. Very often they refer only to one linguistic variable. The theory we have developed can be applied and unifies the study of any alphabetical language.

Of all types of translation, literary translation is maybe the most demanding and difficult because the language of literature is different from ordinary or technical language and involves many challenges on the syntactic, lexical, semantic and stylistic levels [15–22].

Often, in translation theory studies, especially of literary texts, scholars mention the concept of "equivalence" – or "sameness" - between the source text (to be translated) and the target text (translated text) [23–25]. This concept, however is loosely and never mathematically defined. On the

contrary, our theory can mathematically "measure" how much texts differ from each other, with a multidimensional analysis [26], as we do also in the present paper.

Today, it is agreed that the translator of literary texts is an active role player and the intermediary between the source text and the target text [17,27].

In relation to the translator's role, however, two main theories of translation are today discussed by scholars: "domestication" and "foreignization" [27]. Domestication is a translation method in which the translator tries to match the source text to the reader, mainly to reader's reading skills. Foreignization is a method in which the translator tries to match the reader to the source text, essentially regardless of reader's reading skills. In other words, domestication neglects the foreign reality, while foreignization retains the foreignness and cultural otherness of a foreign text.

In translation to English the dominant practice is mainly domestication [27] because very likely translation makes the target text more fluent. This is achieved, however, by erasing the linguistic and cultural difference of the foreign text, therefore, text fluency becomes the general criterion to judge a translation.

In this paper, as study case, we compare by means of a multidimensional mathematical analysis the text of a famous Italian novel, *I promessi sposi* (1840) written by Alessandro Manzoni (1785 – 1873), to the text of the most recent English translation, *The Betrothed* by Michael F. Moore [28], a masterful translation.

To avoid any misunderstanding, the mathematical theory, and its multidimensional application to texts, is not concerned with the efficacy and beauty of a text - in this case the target text - but merely with its mathematical underlying structure, compared, in this case, to the Italian original text. The discussion and conclusion we reach are meant to be valid only in this context. In other words, we try to measure the "equivalence" of the two texts mathematically, just like we have done with other, oftern unrelated, couples of texts in any language, even between different languages [4–8].

Alessandro Manzoni was an Italian poet and novelist. He is famous for the novel *The Betrothed* (In Italian: *I promessi sposi*), ranked among the masterpieces of world literature. He started to write the first version in 1821 after reading *Ivanhoe* by Walter Scott (1771-1832) in a French translation, and revised it in its final version in the years 1830s.

I promessi sposi is a symbol of the Italian *Risorgimento* – the Italian XIX century political renaissance movement which unified Italy as a country – both because of its patriotic message and because the novel was a fundamental milestone in developing the modern Italian language which Manzoni stabilized and ensured its unity throughout Italy. The novel is still today a required reading in Italian High Schools.

Of the two translation theories, domestication and foreignization, Moore clearly has chosen domestication. In fact, in his "Notes on the Translation" he writes [28]:

«Within the framework of fidelity to *I promessi sposi*, I felt adjustments were needed due to fundamental stylistic and grammatical differences between English and Italian conventions. For example, Manzoni often adopts the periodic structure for his sentences, stringing together clause after clause with a series of semicolons. Italian enables and even encourages this style, since as an inflected language, Italian embeds the person and number of the subject—the I, you, or we—in the verb. Subject pronouns are used minimally, so a series of clauses with the same subject can result in a sentence the length of a page. The period, or "full stop," is regarded with barely concealed horror, as if it truncates an otherwise beautiful sentence. English is not as economical in its use of subject pronouns, and actually requires them, so rather than mimic the periodic style at full length, I broke his sentences up into smaller parcels, sometimes (but not always) using periods where he used semicolons, feeling that they do not impose the same sense of finality in English as they do in Italian.»

In the following of the paper, we will realize that the replacement of semicolons with periods has had a very large impact on the underlying mathematical structure of the translation. However, this is not the only change, because other linguistic parameters have been largely modified. The aim of the paper is to show what has been mathematically modified, to what extent and the impact it has on readers.

After this introductory section, Section 2 shows an example of translation (the incipit of the novel), Section 3 reports some total statistics, Section 4 reports an exploratory data analysis of linguistic variables, Section 5 defines deep-language variables, Section 6 defines a geometrical representation of texts and discusses the probability of confusing a text with another, Section 7 deals with the short-term memory of readers/writers, Section 8 deals with the fine-tuning analysis done with linguistic channels, Section 9 deals with a universal readability index and Section 10 summarizes the main results and draw a conclusion.

2. Example of Translation: Incipit

The beginning of *I promessi sposi* – the first sentence of Chapter 1: «Quel ramo del lago di Como…» - is the second most celebrated incipit in the Italian Literature, being Dante's *Inferno* incipit – «Nel mezzo del cammin di nostra vita…» the first one. Generations of Italian High School students have memorized both and still do.

It is illuminating to read how the beginning paragraph was translated (the Italian and English paragraphs, for reader's benefit, are reported in Appendix A). As mentioned in the introduction, here – and in the whole paper - we are not concerned with the meaning and beauty of words, with grammar and so on, but only with the statistics of the linguistic parameters, suitably defined. For this purpose, let us study how this paragraph was translated by calculating some statistics. These arid numbers, however, highlight some important changes done by the translator, which deeply affected the entire English version of the novel. Table 1 reports them.

Table 1. Statistics of the translation of the beginning paragraph of *I Promessi Sposi* (*The Betrothed*).

	Paragraphs	Words	Periods (Sentences)	Commas	Semicolons	Colons
Italian	1	682	9	103	9	5
English	5	701	23	58+8	1	0

The beginning paragraph was split in five. In the Italian paragraph, Manzoni describes Lake Como and its eastern branch, from north to south and down to the main town, Lecco. He describes coves, inlets, hills, mountains, ravines, steep inclines and flat terraces, cultivated fields, vineyards, towns, villages, roads and footpaths. Only when he introduces the first character, Don Abbondio walking along one of the footpaths, Manzoni starts a new paragraph, therefore after 682 words and 9 long sentences with 9 semicolons within them (Table 1).

He seems to act like a modern movie director who sets the stage for the movie first events and ends this introductory paragraph – hence, starts a new paragraph – when in the close-up on a particular footpath the first character appears, Don Abbondio.

Since a paragraph develops a single point, we understand that in Manzoni's mind the single point was the detailed geographical setting of the beginning of his story. When the description is, in his plan, complete, then he starts a new paragraph.

This is not done by the translator. By splitting Manzoni's setting in five paragraphs, he abbandons the original sequence/climax and breaks the flow and unity that Manzoni carefully planned in the entire novel [29–33]. By splitting the single paragraph in five, the translator automatically distinguishes and emphasises five points.

The interpunctions were drastically changed. The most striking change is the almost complete replacement of semicolons and colons with periods. In fact, the sum $9 + 9 + 5 = 23$ gives the number of periods in the translated version. The purpose of semicolons – to separate independent clauses while connecting them as related ideas, just like Manzoni amply does - seems no longer acceptable by modern writers/readers of English texts [34–36]. Of course, this agrees with the translation theory flaunted by the translator (domestication), explicitly stated in his Introduction to

the novel, as recalled in Section 1. Consequently, the liguistic structure of the Italian original text is largely modified, therefore we should expect significant changes in the deep-language parameters.

3. Total Statistics

In this section we report the overall total statistics of linguistic parameters. To calculate them, we have used the digital text (WinWord file) and counted, for each chapter, the number of characters, words, sentences and interpunctions (punctuation marks). Before doing so, we have deleted the titles, footnotes and other extraneous material present in the digital texts. The count is very simple, although time-consuming. Winword directly provides the number of characters and words. The number of sentences was calculated by using WinWord to replace every periods with periods (full stop): of course, this action does not change the text, but it gives the number of these substitutions and therefore the number of periods. The same procedure was repeated for question marks and exclamation marks. The sum of the three totals gives the total number of sentences. The same procedure gives the total number of commas, colons and semicolons. The sum of these latter values with the total number of sentences gives the total number of interpunctions.

Tables 2 and 3 report these total statistics. In English, paragraphs, sentences and words are increased. The characters are fewer than in Italian, but not because the translator decided so, but because English words are, on the average, shorter than Italian words.

The text indicated with Italian-E refers to the Italian text in which $4020 - 263 = 3757$ periods replace semicolons in proportion to the number of words per chapter, to simulate a "reverse" domestication translation. In this case the number of sentences is about the same of the English text, 14676 against 14647, therefore the translator replaced semicolons with periods practically in the entire novel, and this replacement make the number of words per sentence very similar to that of Italian-E text, 16.0 against 15.3. English-I refers to the translation in which semicolons replace periods. In this latter case the number of sentences is smaller than in Italian.

Table 2. Number of paragraphs, words per paragraph, characters contained in the words, words, sentences. Italian refers to the original Manzoni's novel; English refers to Moore's translation; Italian-E refers to the case in which all semicolons of the Italian novel are replace with periods; English-I refers to the translation in which all periods are replace by semicolons (see main text).

	Paragraphs	Words per paragraph	Characters	Characters per word	Words	Words per sentence	Sentences
Italian	2732	82.1	1,036,560	4.62	224,234	21.1	10627
Italian–E	2732	82.1	1,036,560	4.62	224,234	15.3	14647
English	3029	77.5	1,022,239	4.36	234,646	16.0	14676
English-I	3029	77.5	1,022,239	4.36	234,646	30.6	7672

Table 3. Number of periods, question marks, exclamation marks, commas, semicolons, colons and total interpunctions. Italian refers to the original Manzoni's novel; English refers to Moore's translation.

	Periods	Question Marks	Exclamation Marks	Commas	Semicolons	Colons	Interpunctions
Italian	7795	1335	1497	26316	4020	2633	43596
English	11692	1558	1426	20003	263	540	35482

In the next section we report an exploratory data analysis of the two texts. with the aim of visualing important relations between linguistic variables and further investigate the changes done by the translator.

4. Exploratory Data Analysis: Relationships Between Italian and English Linguistic Variables

This section compares the linguistic variables per chapter of the English and Italian texts, 39 samples due to the Introduction (the one written by Manzoni) and 38 chapters of the novel. These scatterplots will be further studied in Section 8 according to the theory of linguistic channels.

Figure 1(a) shows the difference between the number of paragraphs per chapter in the two texts, $n_{par,Eng} - n_{par,Ita}$ and the percentage change $100 \times (n_{par,Eng} - n_{par,Ita})/n_{par,Ita}$. We notice that the translator split the Italian paragraphs by increasing them up to 60%. Figure 1(b) shows the scatterplot between the number of paragraphs; it clearly shows that the data are all above the 45° line $y = x$. Table 4 reports the correlation coefficient and slope of the regression line for this scatterplot and also for the following ones.
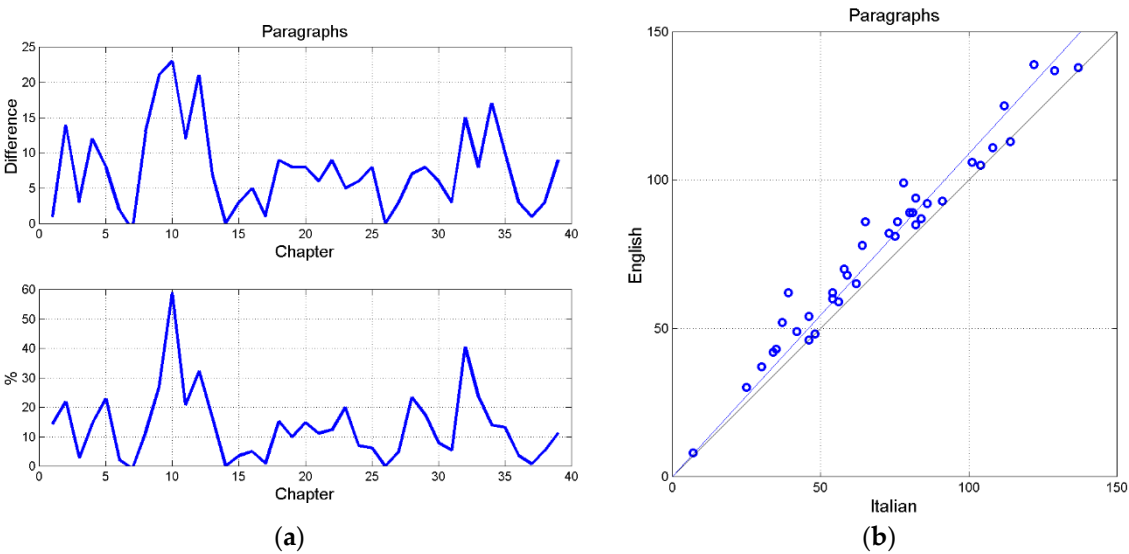

Figure 1. (a) Difference and percentage change of paragraphs per chapter in the two texts; (b) Scatterplot between the two variables. The sample size is 39.

Table 4. Correlation coefficient and slope of the scatterplots between Italian and English variables (samples refer to chapeters) for which a linear relationship fits the data.

Linguistic Variable	Correlation Coefficient	Slope
Characters	0.9973	0.9861
Paragraphs	0.9809	1.0897
Words	0.9967	1.0470
Sentences	0.9771	1.3704

Sentences (Italian-E)	0.9868	1.0047
Question Marks	0.9820	1.1581
Exclamation Marks	0.9237	0.9215
Commas	0.9477	0.7572
Interpunctions	0.9826	0.8143

Figurea 2(a) shows the scatterplot between characters and Figure 2(b) shows that between words. Both variables are very much correlated. In English the number of words is always larger than in Italian, therefore the translator needs more words to convey the original meaning. Characters and words are the most correlated variables.
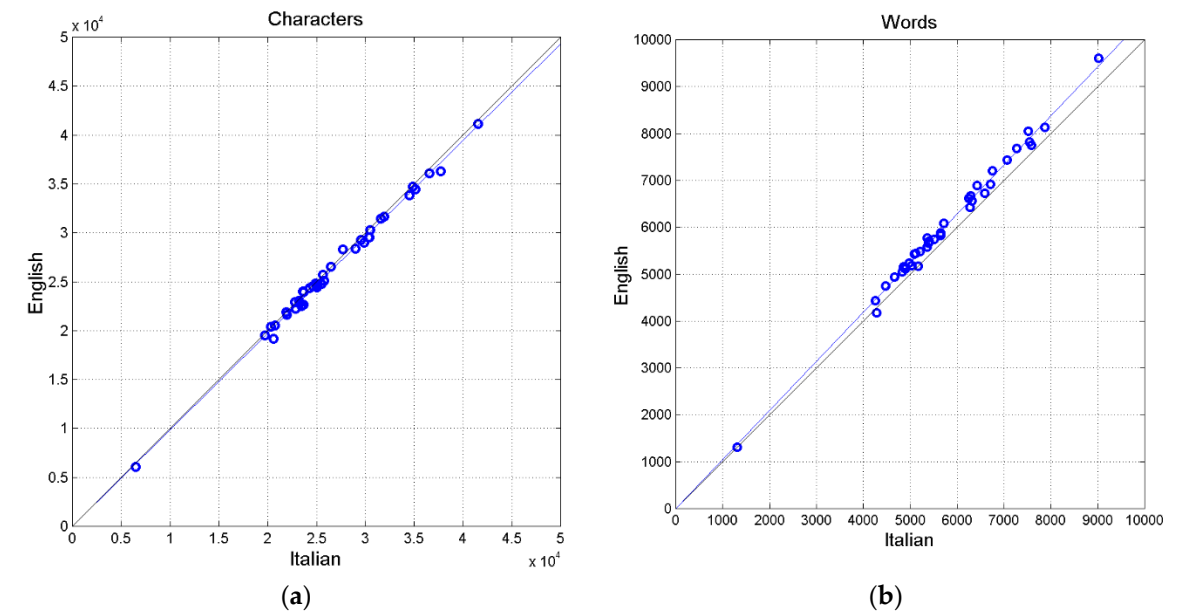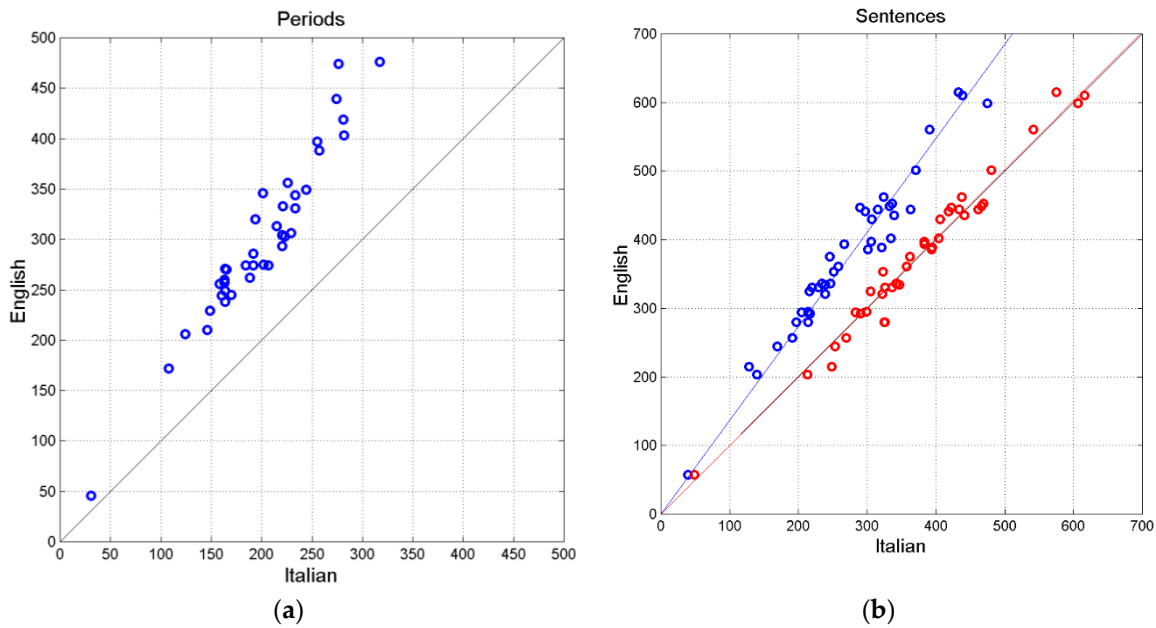

Figure 2. (a) Scatterplot between characters per chapter; (b) Scatterplot between words per chapter.
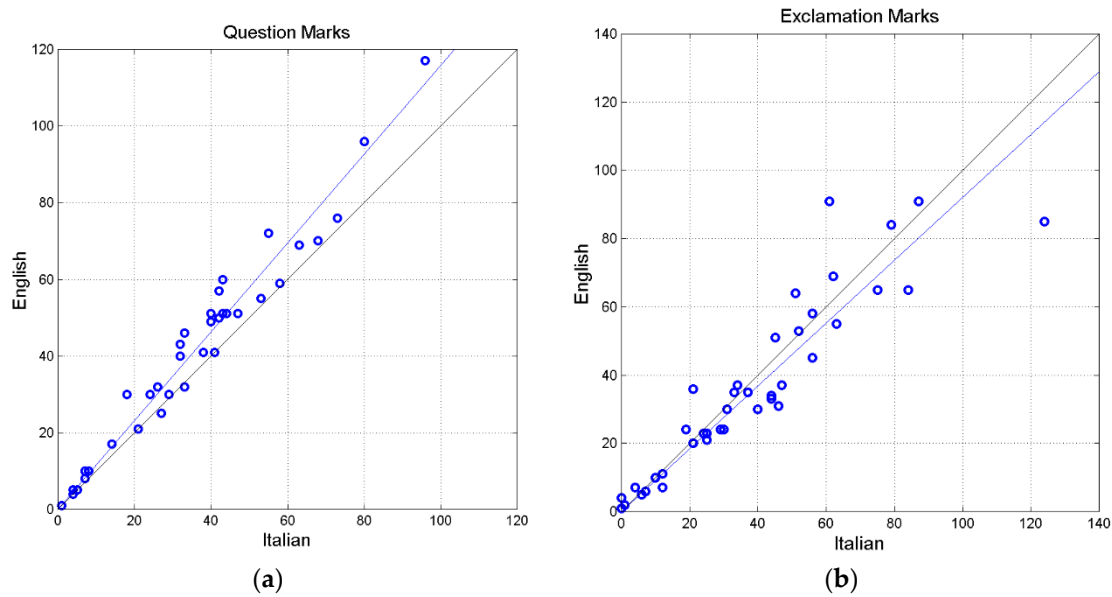
Figure 3(a) shows the scatterplot between periods. The replacement of most semicolons with periods is clearly evidenced by being all samples above the 45° line $y = x$.

Figure 3(b) shows the scatterplot between sentences. The effect of replacing semicolons with periods is clearly evident in the blue circles and regression line. Since the Italian text contains many semicolons, sentences are quite long, therefore the regression line is far from the $y = x$ line. On the contrary, when in Italian the semicolons are counted as periods, then the agreement is very good, as the red circles and regression line show.

Figures 4(a) shows the scatterplot between question marks and Figure 4(b) between exclamation marks. It is surprising that both do not follow at all the Italian text (i.e. the 45° line), particularly the number of questions marks, which in English is larger than in Italian. Appendix C reports a short example in which question and exclamation marks are modified.
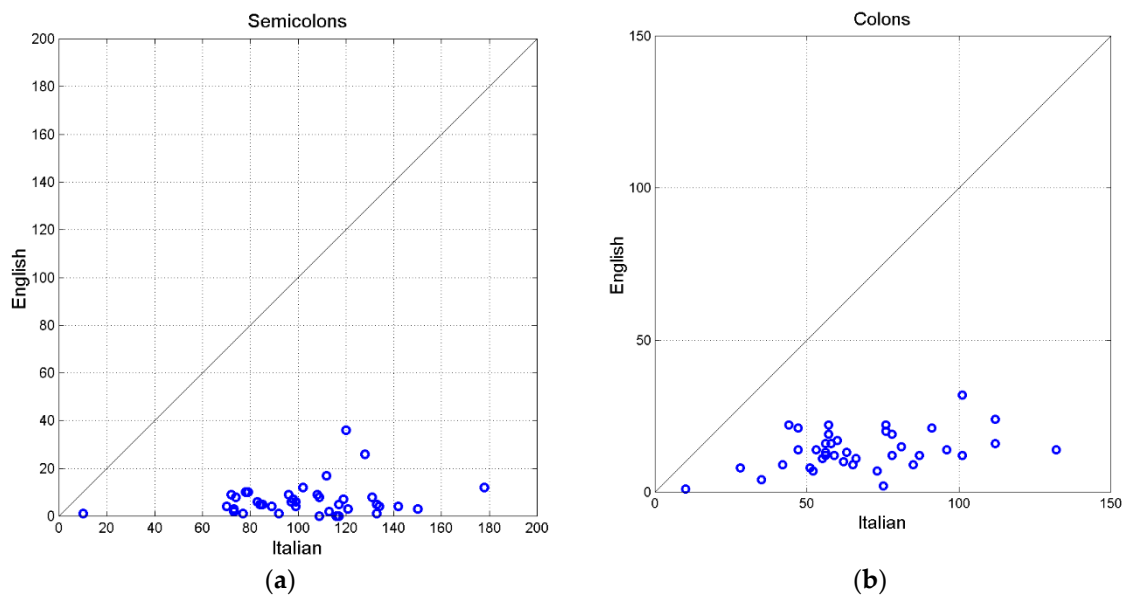
Figure 3. (a) Scatterplot between periods per chapter; (b) Scatterplot between sentences per chapter. Blue indicates the comparison with the originall Italian text; red indicates the comparison with the Italian text in which semicolons are replaced by periods.

Figure 4. (a) Scatterplot between question marks per chapter; (b) Scatterplot between exclamation marks per chapter.

Figure 5(a) shows the scatterplot between semicolons and Figure 5(b) shows that between colons. The almost total cancellation in English of both interpunctions is clearly evident.

Figure 5. (a) Scatterplot between semicolons per chapter; (b) Scatterplot between exclamation marks per chapter.

Figure 6(a) shows the scatterplot between commass. We notice that in English there are quite fewer commas than in Italian and this reduction, together with that in colons, semicolons, question and exclamation marks make the number of interpunctions quite smaller than in Italian, Figure 6(b).
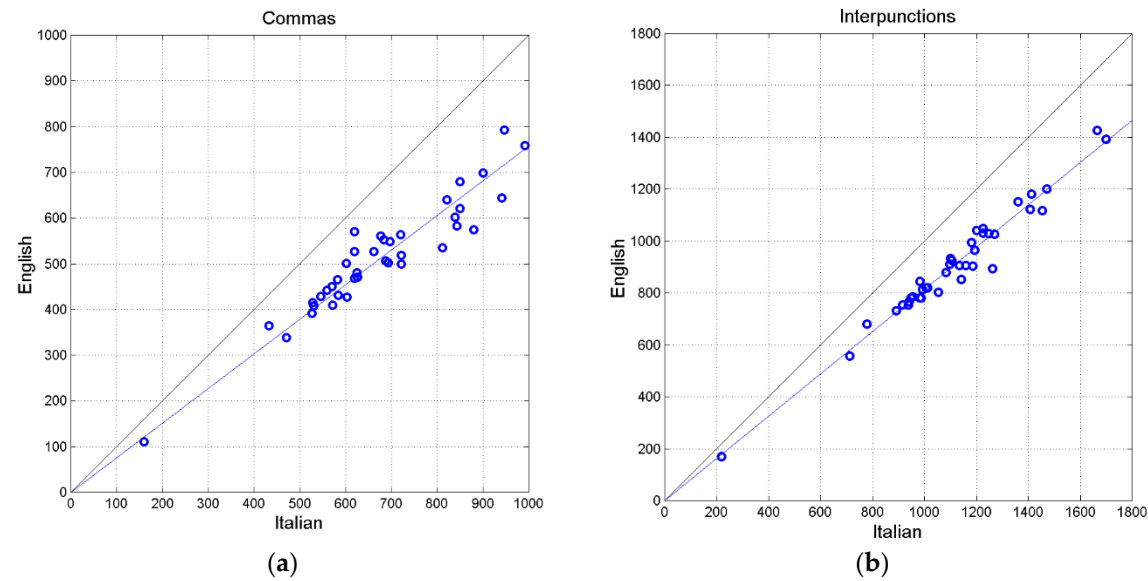
Figure 6. (a) Scatterplot between commas per chapter; (b) Scatterplot between total interpunctions.

In conclusion, the translator completely modified the underlying texture of interpunctions by distributing them quite differently from the original Italian text, therefore changing reader's pauses and reader's short-term memory load, as discussed in Section 7 below. A similar modification occurs also in the distribution of paragraphs.

Only characters, Fiogure 2(a), follow very strictly the 45° line, but this large agreement is not a translator's choice but very likely due to two remarks: (a) both Italian and English are alphabetical languages with many similar words coming directly from Latin in Italian and indirectly, through French, in English; (b) alphabetical languages have very similar mean number of characters per word (see Table 1 of Ref.[37]).

In other terms, a blind comparison of the statistics concerning fundamental linguistic parameters would unlikely indicate these two texts refer to the same novel, although written in different languages.

In the next section, we recall the deep-language variables which allow further studies.

5. Deep–Language Variables

Let us recall four deep-language variables [1,2] which are not consciously managed by writers, therefore useful for assessing similarity of texts beyond writers' awareness. To avoid possible misunderstanding, these variables refer to the "surface" structure of texts, not to the "deep" structure mentioned in cognitive theory.

Let n_C, n_W and n_I be respectively the number of characters, words and interpunctions (punctuation marks) calculated in disjoint blocks of texts, such as chapters (in this paper) or any other subdivisions, then we can define four deep–language variables (Appendix B lists the mathematical symbols used in the present paper):

The number of characters per word, C_P:

$$C_P = \frac{n_C}{n_W} \tag{1}$$

The number of words per sentence, P_F:

$$P_F = \frac{n_W}{n_S} \tag{2}$$

The number of interpunctions per word, referred to as the word interval, I_P:

$$I_P = \frac{n_I}{n_W} \tag{3}$$

The number of word intervals, n_{I_P}, per sentence, M_F:

$$M_F = \frac{n_{I_P}}{n_S} \tag{4}$$

Table 5 reports mean and standard deviation of these variables. These values have been calculated by weighing each text with its number of words to avoid that shorter texts weigh statistically as much as long ones. In other words, the statistical weight of a chapter is given by the ratio between the number of its words and the total number of words of the novel. Moreover, notice that the average values of these variables calculated from the totals, as done in Table 2 do not coincide with the arithmetic or statistical means (the ones reported in Table 5), as is proved in Appendix D.

From these statistical mean values, we notice:

a) $< C_p >$ is very similar in both languages, for the reason recalled in Section 4.

b) $< P_F >$ in Italian is, as expected, quite larger than in English. It becomes smaller and very similar to that in English only if semicolons are replaced by periods (Italian-E).

c) $< I_P >$ is significantly smaller in Italian than in English, due to the large number of interpunctions present in Italian. This parameter does not depend on the type of interpunctions, therefore it is the same also in Italian-E.

Table 5. Mean values and standard deviation of the indicated deep–language variables. Italian refers to the original Manzoni's novel; English refers to Moore's translation; Italian-E refers to the case in which all semicolons of the Italian novel are replaced by periods; English-I refers to the translation in which the number of periods are replaced by semicolons.

	$< C_p >$	$< P_F >$	$< I_P >$	$< M_F >$
Italian	4.62	22.7	5.18	4.34
	0.12	6.24	0.41	1.03
Italian–E	4.62	16.01	5.18	3.07
	0.12	--	0.41	--
English	4.36	16.82	6.66	2.50
	0.15	4.56	0.56	0.43
English-I	4.36	30.30	6.66	4.53
	0.15	--	0.56	--

The mean values of Table 5 - or their minimum values directly calculated from the totals, as discussed in Appendix D - are useful for a first assessment of "similarity"of texts by defining linear combinations of deep–language parameters and then modeling them as vectors, a representation briefly recalled in the next section.

6. Geometrical Representation of Alphabetical Texts

The mean values of Table 5 can be used to model texts as vectors, representation discussed in detail in Refs.[1–3] and here briefly recalled. This geometrical representation allows also to calculate the probability that a text/author can be confused with another one, an extension in two dimensions of the problem discussed in Ref. [6]. The conditional probability can be considered an index indicating who influenced who, as shown in Ref.[26]. In our case it gives another indication on how much the translation differs from the original text.

6.1. Vector Representation of Texts

Let us consider the following six vectors of the indicated components of deep–language variables: $\vec{R_1} = (< C_P >, < P_F >)$, $\vec{R_2} = (< M_F >, < P_F >)$, $\vec{R_3} = (< I_P >, < P_F >)$, $\vec{R_4} = (< C_P >, < M_F >)$, $\vec{R_5} = (< I_P >, < M_F >)$, $\vec{R_6} = (< I_P >, < C_P >)$ and their resulting vector sum:

$$\vec{R} = \sum_{k=1}^{6} \vec{R_k} = x\vec{i} + y\vec{j} \qquad (5)$$

The choice of which parameter represents the component in abscissa and ordinate axes is not important because, once the choice is made, the numerical results will depend on it, but not the relative comparisons and general conclusions.

In the first quadrant of the Cartesian coordinate plane two texts are likely mathematically connected – they show close ending points of vector (5) – if their relative Pythagorean distance is small.

To set the two texts here studied - especially the translated text - in the framework of the English Literature, we have considered the same Cartesian plane defined and discussed in Ref. [26], which reports the list of the English novels whose positions are reported in this figure.

By considering the vector components x and y of Eq. (5), we obtain the scatterplot shown in Figure 7 where X and Y are normalized coordinates calculated by setting *Lord of the Rings* at the origin ($X = 0$, $Y = 0$) and *Silmarillion* at ($X = 1, Y = 1$), according to the linear tranformations:

$$X = \frac{x - x_{Lord}}{x_{Silma} - x_{Lord}} \qquad (6)$$

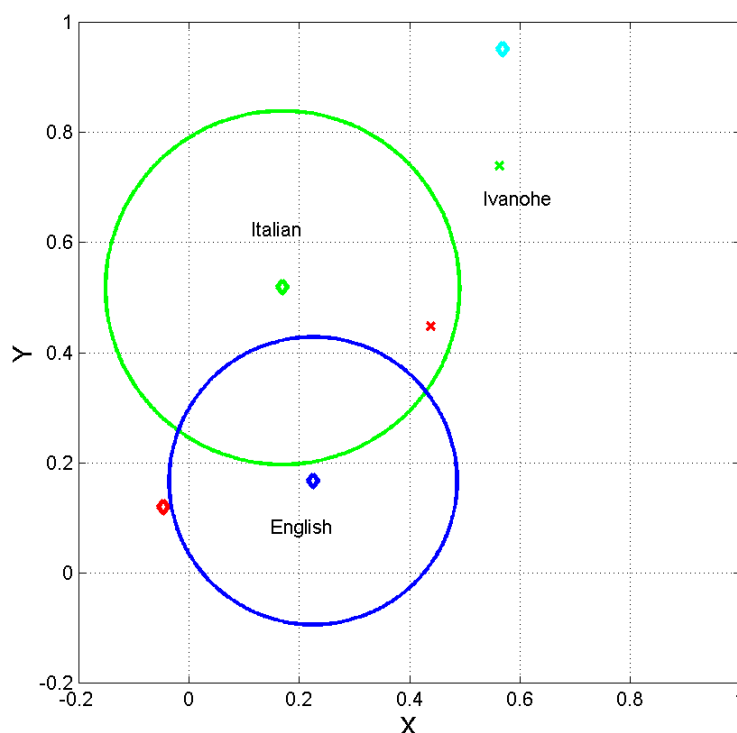$$X = \frac{y - y_{Lord}}{y_{Silma} - y_{Lord}} \qquad (7)$$

In Figure 7 we report the position of the Italian text (green diamond), its translation (blue diamond) and the position of *Ivanohe* - a novel that inspired Manzoni in writing his novel - with green ×. Notice that since $< C_P >$ in the two languages is very similar (see Table 5), the position depends mainly on the other mean values.

We can notice that the translation (blue diamond) is well within the other English novels, in particular it almost concides with *The Jungle Book*. The Italian text is significantly displaced from most English novels.

Figure 8 shows a zoom. We can notice that if periods are replaced by semicolons in the translation (English-I, cyan diamond), then this text is located very near *Ivanohe*, which may indicate an indirect influence of Walter Scott on Manzoni's writing. On the other hand, if semicolons are replaced by periods, *Ivanohe* gets closer to English (red " × "). The red diamond refers to Italian-E, which falls well within the English texts reported in Figure 7.
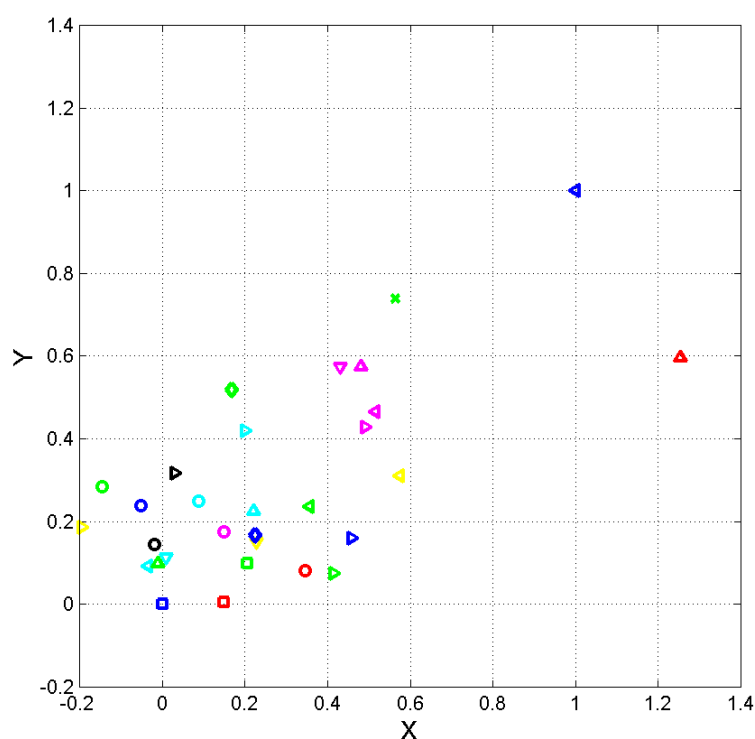
Figure 7. Normalized coordinates X and Y of the ending point of vector (5) such that *The Lord of the Rings*, blue square, is at (0,0) and *The Silmarillion*, blue triangle pointing left, is (1,1). *The Chronicles of Narnia*: red square; *The Space Trilogy*: red circle; *The Hobbit*: blue triangle pointing right; *The Screwtape Letters*: red triangle pointing upward; *At the Back of the North Wind*: cyan triangle pointing left; *Lilith: A Romance*: cyan triangle pointing downward; *Phantastes: A Fairie Romance for Men and Women*: cyan triangle pointing right; *The Little Princess*: cyan triangle pointing upward; *Oliver Twist*: blue circle; *David Copperfield*: green circle; *A Tale of Two Cities*: cyan circle; *Bleak House*: magenta circle; *Our Mutual Friend*: black circle; *Pride and Prejudice*: magenta triangle pointing right; *Vanity Fair*: magenta triangle pointing left; *Moby Dick*: magenta triangle pointing downward; *The Mill On The Floss*: magenta triangle pointing upward; *Alice's Adventures in Wonderland*: yellow triangle pointing right; *The Jungle Book*: yellow triangle pointing downward; *The War of the Worlds*: yellow triangle pointing right; *The Wonderful Wizard of Oz*: green triangle pointing left; *The Hound of The Baskervilles*: green triangle pointing right; *Peter Pan*: green triangle pointing upward; *Martin Eden*: green square; *Adventures of Huckleberry Finn*: black triangle pointing right. See [] for the full details on the novels. Texts considered in the present paper: *I promessi sposi*: green diamond; *The betrothed*: blue diamond; *Ivanohe*: green " × ".

Figure 8. Normalized coordinates X and Y of the ending point of vector (5) such that *Lord of the Rings* is at (0,0) and *Silmarillion* is at (1,1), zoom. Italian text: green diamond; its translation: blue diamond; Italian text with semicolons transformed in periods (Italian-E: red diamond; English text with periods transformed in semicolons (English-I: cyan diamond. The position of *Ivanohe* is reported with green " × " (original text) and with red " × " if semicolons are replaced by periods. The green and blue circles represent 1-sigma contour lines discussed in the main text.

These remarks refer to the display of vectors whose ending point depends only on mean values. The standard deviation of the four deep−language variables, reported in Table 5, do introduce data scattering, therefore in the next sub−section we study and discuss this issue by calculating the probability (called "error" probability) that a text may be mathematically confused with another one, as done in Ref. [26].

6.2. Error Probability

Besides the vector $\vec{R}$ of Eq. (5) - due to mean values - we can consider another vector $\vec{\rho}$, due to the standard deviation of the four deep−language variables, that adds to $\vec{R}$ [26]. In this case, the final random vector describing a text is given by:

$$\vec{T} = \vec{R} + \vec{\rho} \qquad (8)$$

Now, to get some insight into this new description, we consider the area of a circle centered at the ending point of $\vec{R}$.

We fix the magnitude (radius) ρ as follows. First, we add the variances of the deep−anguage variables that determine the components x and y of $\vec{R}$ − let them be σ_x^2, σ_y^2 − then we calculate the average value $\sigma_\rho^2 = 0.5 \times (\sigma_x^2 + \sigma_y^2)$ and finally we set:

$$\rho = \sigma_\rho \qquad (9)$$

Now, since in calculating the coordinates x and y of $\vec{R}$ a deep−language variable can be summed twice or more, we add its standard deviation (referred to as sigma) twice or more times before squaring, as shown in [26].

Now, we can estimate the (conditional) probability that a text is confused with another by calculating ratios of areas. This procedure is correct if we assume that the bivariate density of the normalized coordinates ρ_X, ρ_Y, centred at $\vec{R}$, is uniform. By assuming this hypothesis, we can

calculate probabilities as ratios of areas [38, 39]. As dicussed in [26], the hypothesis of substantial uniformity around $\vec{R}$ should be justified within 1-sigma bounds.

Now we can calculate the following probabilities. Let A be the common area of two 1−sigma circles (i.e., the area proportional to the joint probability of two texts), let A_1 be the area of 1−sigma circle of text 1 and A_2 the area of 1−sigma circle of text 2. Now, since probabilities are proportional to areas, we get the following relationships:

$$\frac{A}{A_1} = \frac{P(A_1, A_2)}{P(A_1)} = \frac{P(A_2/A_1)P(A_1)}{P(A_1)} = P(A_2/A_1) \tag{10}$$

$$\frac{A}{A_2} = \frac{P(A_1, A_2)}{P(A_2)} = \frac{P(A_1/A_2)P(A_2)}{P(A_2)} = P(A_1/A_2) \tag{11}$$

Therefore, A/A_1 gives the conditional probability $P(A_2/A_1)$ that part of text 2 can be confused (or "contained") in text 1; A/A_2 gives the conditional probability $P(A_1/A_2)$ that part of text 1 can be confused with text 2. $P(A_2/A_1) = 1$ means [38,39 $A = A_1$, therefore text 1 can be fully confused with text 2, and $P(A_1/A_2) = 1$ means $A = A_2$, therefore text 2 can be fully confused with text 1.

Now, we can recall a synthetic parameter which highlights how much two texts can be erroneously confused with each other. The parameter is the average conditional probability:

$$p_e = P(A_2/A_1)P(A_1) + P(A_1/A_2)P(A_2) \tag{12}$$

Now, since in comparing two texts we can assume $P(A_1) = P(A_2) = 0.5$, we get:

$$p_e = 0.5 \times [P(A_2/A_1) + P(A_1/A_2)] \tag{13}$$

If $p_e = 0$, there is no intersection between the two 1−sigma circles, the two texts cannot be each other confused, therefore there is no mathematical connection involving the deep−language parameters. If $p_e = 1$ the two texts can be totally confused, the two 1−sigma circles coincide.

In our exercise, assuming the Italian text as text 1 and its translation as text 2, we get $P(A_2/A_1) = 0.225$ and $P(A_1/A_2) = 0.339$, therefore $p_e = 0.282$. Since this value is very low, there is little chance that the two texts can be each other confused. In conclusion, as already remarked, a blind analysis of the two texts will unlikely reveal their real relationship.

In the next section we study and discuss how the Italian text and its translation engage readers' short-term memory.

7. Short−Term Memory of Writers/Readers

The human short−term memory seems to be modelled by two practically independent variables, which apparently engage two short−term memory (STM) buffers in series [7,8]. The first buffer is modelled according to the number of words between two consecutive interpunctions - i.e., the variable I_P, the word interval -which follows Miller's 7 ± 2 law [40] (on the large literature on this topic, see the studies and remarks reported in [41–63]). The second buffer is modelled according to the number of word intervals, the variable M_F, contained in a sentence, referred to as the extended short-term memory (E-STM).

As for the first buffer, Figure 9 shows I_P versus P_F. Notice that the replacement of semicolons with periods is clearly evident (red circles). The trends shown are typical of alphabetical languages [1–8,26] and confirm two fundamental aspects:

a) I_P ranges approximately in Miller's bounds.

b) As the number of words in a sentence, P_F, increases, I_P can increase but not linearly, because the first buffer cannot hold, approximately, a number of words larger than that empirically predicted by Miller's Law, therefore saturation *must* occur. Scatterplots like that shown in Figure 9 give an insight into the short−term memory capacity engaged in reading/writing a text, because a writer is also a reader of his/her own text.
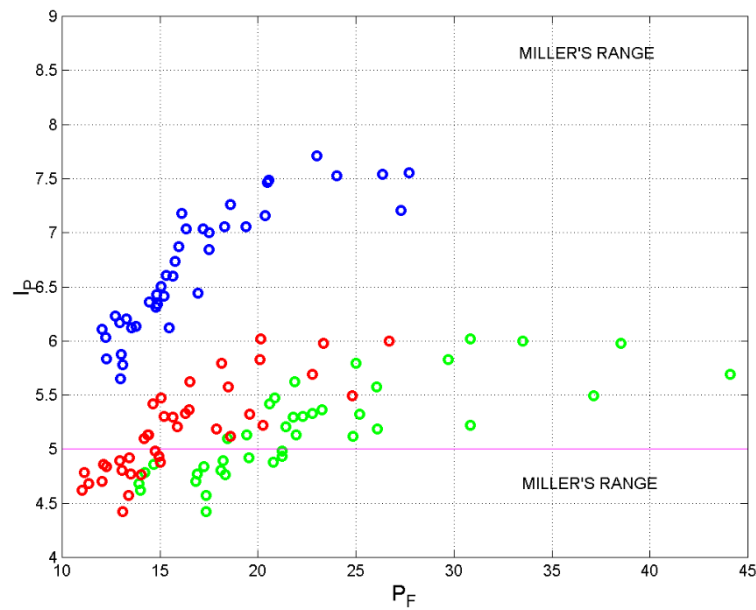
Figure 9. I_P versus P_F. Italian: green circles; English: blue circles; Italian-E: red circles. Miller's range is between $I_P = 5$ and $I_P = 9$.

As for the second buffer, Figure 10 shows the scatterplot between M_F and I_P. The effect of replacing semicolons with periods is clearly evident (red circles). Recall that I_P is insensitive to this replacement because n_I does not change.

More interestingly, in translation $M_F \approx 2.5 = < M_F >$ (see Table 5) for a large range, while in Italian it varies more randomly. In other words, the effect of replacing semicolons with periods (English) has the effect of forcing the capacity of the second buffer to a constant value. Finally, Figure 10 shows that, for a large range I_P and M_F are uncorrelated, therefore independent, according to the log-normal probability modelling discussed in Ref. [26]. In other words, also the reading of these specific texts seem to require a STM made of two independent buffers in series.
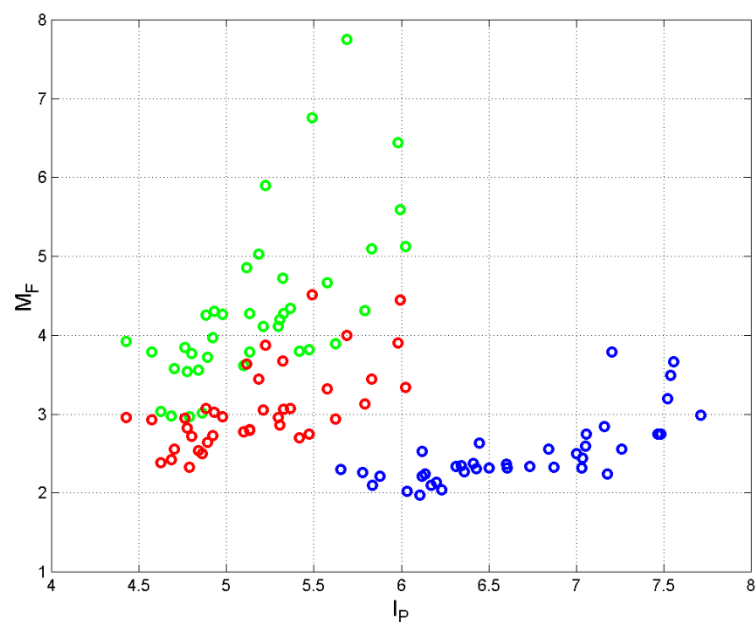
Figure 10. M_F versus I_F. Italian: green circles; English: blue circles; Italian-E: red circles.

In [8] we have studied the numerical patterns (due to combinations of I_P and M_F) which give the number of sentences that theoretically can be recorded in the two STM buffers, regardless of their meaning. This number was compared with that of the sentences actually found in novels of Italian and English Literatures and found that most authors write for readers with small STM buffers and, consequently, are forced to reuse numerical patterns to convey multiple meanings.

Figure 11 shows the scatterplot between the theoretical number of sentences allowed in Italian and in English. In Italian this number is about two order of magnitude greater than in English, a find mainly due to the larger M_F (see Figure 10). Only when semicolons are replaced by periods in the Italian text (red circles) the two variables agree more. In any case, in Italian the number of possible patterns is always larger than in English.
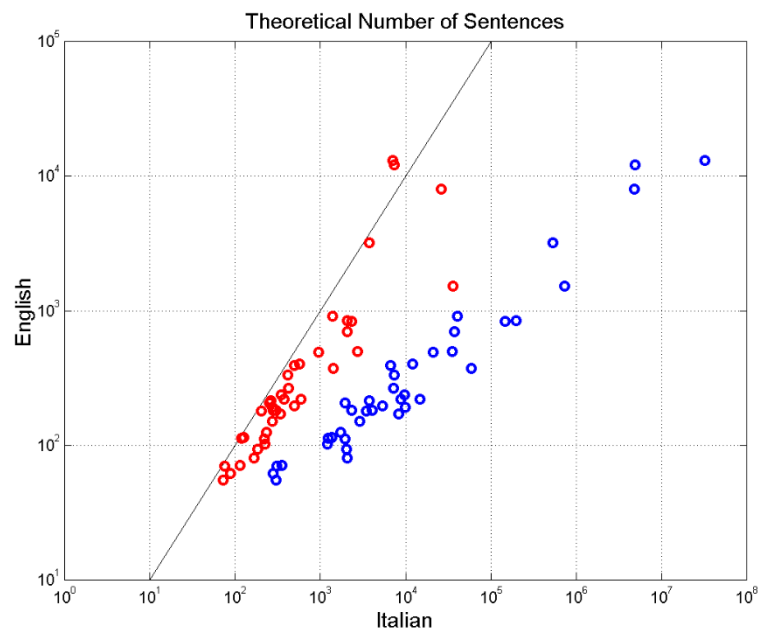
Figure 11. Scatterplot between the theoretical number of sentences per chapter allowed in Italian and that in English: blue circles. Scatterplot between Italian-E and English: red circles.

The ratio between the number of sentences in a novel and the number of sentences theoretically allowed by the two buffers concerning I_P and M_F, referrred to as the multiplicity factor α, can be used to compare two texts [8]. In the Italian and English Literatures we found that $\alpha > 1$ is more likely than $\alpha < 1$, and often $\alpha \gg 1$. In the latter case, writers reuse many times the same pattern of sentences.

In our present case, we study a text and its translation. Figure 12 shows α for the two texts. Differences are clearly evident and they depend, of course, on the finds reported in Figure 11, with very low values in Italian. Italian and English are clearly very different. In other words, Manzoni has potentially many patterns to use, according mainly to M_F, but he does not. These patterns are much reduced in English; only Italian-E (red circles) agrees more with English.

Another useful parameter is the mismatch index [8], which measures to what extent a writer uses the number of sentences theoretically available, defined by:

$$I_M = \frac{\alpha - 1}{\alpha + 1} \qquad (14)$$

$I_M = 0$ if $\alpha = 1$; in this case experimental and theoretical patterns are perfectly matched. They are over matched if $I_M > 0$ ($\alpha > 1$) and under matched if $I_M < 0$ ($\alpha < 1$).

Figure 13 shows the scatterplot of I_M between the two texts. Again there is a large difference between Italian and English (blue circes). Only Italian-E agrees with English (red circles).
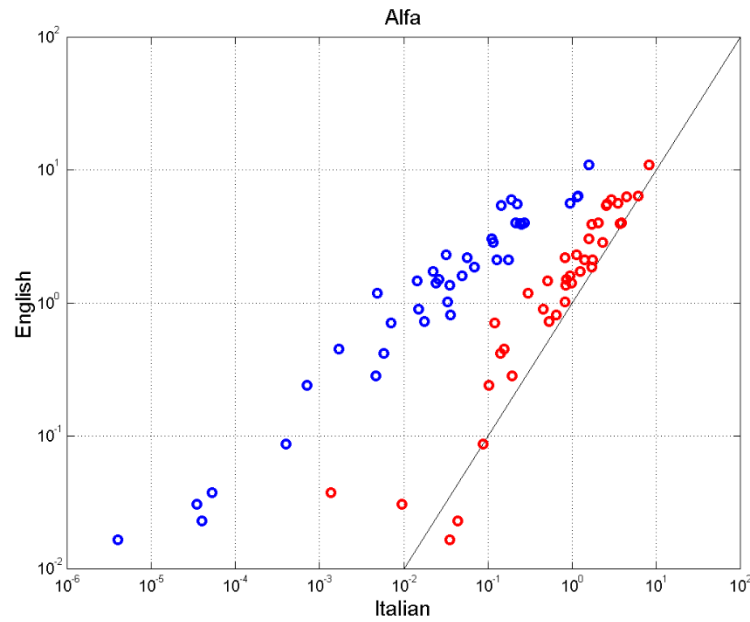
Figure 12. Scatterplot between the multiplicity factor α in Italian and in English: blue circles. Scatterplot between Italian-E and in English: red circles.
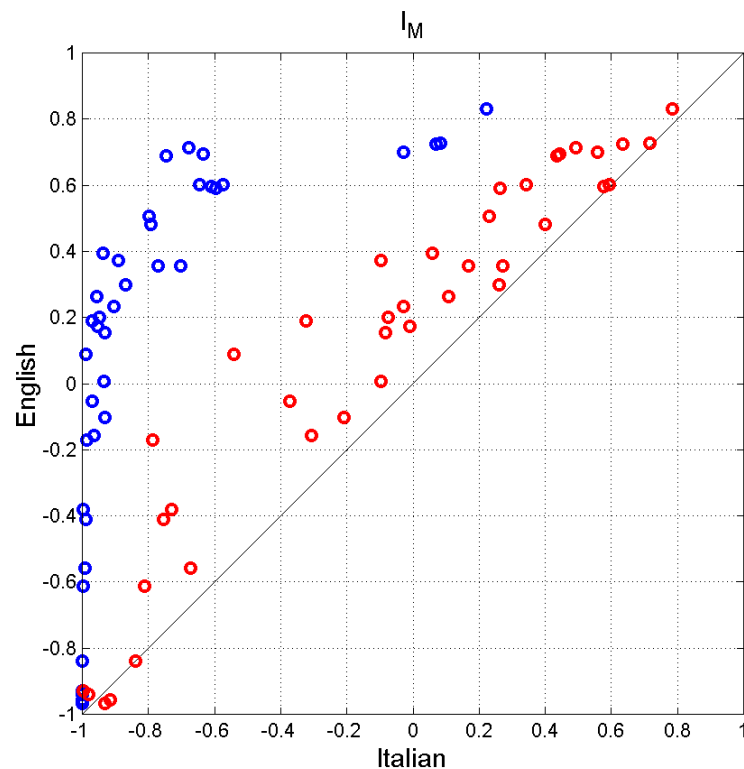
Figure 13. Scatterplot between the mismatch factor I_M in Italian and in English: blue circles. Scatterplot between Italian-E and in English: red circles.

In conclusion, in translation the STM reader's buffers are so much diverse to make the two texts hardly recognizable as one the translation of the other. The two texts show significant correlation only if semicolons are replaced by periods in the Italian text, which indicates, as already noticed, that almost all semicolons were replaced with periods.

The next section recalls the theory of important linguistic channels.

8. Linguistic Channels

In this section, we first recall important linguistic channels present in alphabetical texts, which represent a "fine-tuning" analysis of linguistic variables; secondly the general theory of linear channels [3,6] concerning the joint processing of two experimental scatterplots; thirdly the theory applied to the particular case of processing a single scatterplot, and finally the application of the thoery to the two texts.

Notice that since the theory deals with linear regressions, it can be applied and be useful in any field in which linear relationhsips fit experimental data.

8.1. Linguistic Channels

In alpahabetical texts we can always define at least four linguistic linear channels [3,6], namely:

(a). Sentence channel (S–channel)
(b). Interpunctions channel (I–channel)
(c). Word interval channel (WI–channel)
(d). Characters channel (C–channel).

In S–channels, the number of sentences of two texts is compared for the *same* number of words. From the two scatterplots relating n_S to n_W in two texts, we can relate $n_{S,j}$ of text Y_j to $n_{S,k}$ of text Y_k by setting $n_{W,j} = n_{W,k}$. These channels describe how many sentences the author of text j writes, compared to the writer of text k (reference text), by using the same number of words. Only the theory of linguistic channels allows this comparison. These channels are more linked to P_F than to the other variables. Very likely they reflect the style of the writer.

In I–channels, the number of word intervals, M_F, of two texts is compared for the *same* number of sentences. These channels describe how many short texts between two contiguous punctuation marks (of length I_P words) two authors use. Since M_F is connected to the E–STM, I–channels are more related to the second buffer capacity than to the style of the writer.

In WI–channels, the number of words contained in a word interval, (i.e., I_P) is compared for the *same* number of interpunctions, therefore WI–channels are more related to the first buffer of the STM than to the style of the writer.

In C–channels, the number of characters of two texts is compared for the same number of words. These channels are more related to the language used than to other variables.

8.2. General Theory of Linear Channels

In a scatterplot involving – in our case - linguistic variables, an independent (reference) variable x and a dependent variable y can be related by a regression line (slope m) passing through the origin:

$$y = mx \qquad (15)$$

However, Eq. (15) does not give the full relationship between x and y because it links only mean conditional values. For example, for two diverse texts Y_k and Y_j, we can write Eq. (15) for the same variables x and y.

More general linear relationships can be written by considering also the scattering of the data—measured in Y_k and Y_j by the correlation coefficients r_k and r_j, not considered in Equation (12)—around the regression lines (slopes m_k and m_j) in two scatterplots:

$$y_k = m_k x + n_k \qquad (16)$$

$$y_j = m_j x + n_j$$

While Eq. (15) connects the dependent variable y to the independent variable x only on the average, Eq. (16) introduces additive "noise" n_k and n_j, with zero mean value. The noise, therefore, is due to the correlation coefficient $|r| \neq 1$.

The theory allows to compare two dependent variables y_j and y_k by eliminating x. We can calculate the slope and correlation coefficient of y_j versus y_k without knowing their scatterplot. For example, we can compare the number of sentences in two texts for an equal number of words by considering not only the two regression lines that link sentences to words – which would provide only the slope of the linear relationship between y_j and y_k - but also the scattering of the data. In this particular case, we refer to this linguistic channel as the "sentences channel" and to the channel processing as "fine tuning" because it deepens the analysis of the data and provides more insight into the relationship between two texts.

The mathematical theory is here briefly recalled (more details in Refs. [2,3]). Continuing the example of sentences, we can study the "sentences channel" that relates the sentences of the input text Y_k to those of the output text Y_j by eliminating x (i.e., the variable words). From Eq. (16) we obtain the linear noisy relationship between the sentences in text Y_k (reference, independent, channel input text) and the sentences in text Y_j (dependent, channel output text):

$$y_j = \frac{m_j}{m_k} y_k - \frac{m_j}{m_k} n_k + n_j \qquad (17)$$

In Eq.(17) the slope of the regression line linking y_j to k, m_{jk}, is given by:

$$m_{jk} = \frac{m_j}{m_k} \qquad (18)$$

The "regression noise–to–signal ratio", R_m, due to $|m_{jk}| \neq 1$, of the channel is given by:

$$R_m = (m_{jk} - 1)^2 \qquad (19)$$

The unknown correlation coefficient r_{jk} between y_j and y_k is given by [39]:

$$r_{jk} = \cos\left|\text{arcos}(r_j) - \text{arcos}(r_k)\right| \qquad (20)$$

The "correlation noise–to–signal ratio", R_r, due to $|r_{jk}| < 1$, is given by:

$$R_r = \frac{1 - r_{jk}^2}{r_{jk}^2} m_{jk}^2 \qquad (21)$$

Because the two noise sources are disjoint, the total noise–to–signal ratio R of the channel connecting text Y_k to text Y_j is given by:

$$R = (m_{jk} - 1)^2 + \frac{1 - r_{jk}^2}{r_{jk}^2} m_{jk}^2 \qquad (22)$$

From Eq.(22) we obtain the signal-to-noise ratio γ (natural units) and Γ (dB)

$$\gamma = \frac{1}{R} \qquad (23)$$

$$\Gamma = 10 \times \log_{10}\gamma \qquad (24)$$

Notice that no channel can yield $|r_{jk}| = 1$ and $|m_{jk}| = 1$ (i.e., $\gamma = \infty$), a case referred to as the ideal channel, unless a text is compared with itself (self–comparison, self–channel). In practice, we always find $|r_{jk}| < 1$ and $|m_{jk}| \neq 1$.

In summary, the slope m_{jk} measures the multiplicative "bias" of the dependent variable compared to the independent variable; the correlation coefficient r_{jk} measures how "precise" the linear best fit is. The slope m_{jk} is the source of the "regression noise" of the channel, the correlation coefficient r_{jk} is the source of the "correlation noise".

Finally, notice that since the probability of jointly finding $|r| = 1$ and $m = 1$ is practically zero, all channels are always noisy.

8.3. The Channel with a Single Scatterplot: One–to–One Correspondence

To clarify what we mean by a single scatterplot and one–to–one correspondence, let us consider the translation of a text in which we draw a scatterplot for each linguistic variable. For example, we can display the number of words per chapter in the translated text versus that of the original text, as done in Figure 1 in which $m = 1.047$ and $r = 0.9967$. Now, if there were a perfect ideal translation, the translated text would have the same number of words per chapter of the original text, therefore $m = 1$ and $r = 1$ because the scatterplot would collapse to the deterministic $45°$ linear relationship between x and y, giving the reelationship:

$$y_k = x \qquad (25)$$

Let us show how the relationships of Section 8.2 change in this particular and ideal case. If $m_k = 1$ and $n_k = 0$, then $r_k = 1$, therefore this is a special case of the general theory recalled in the previous sub-section, therefore we get the following expressions:

$$m_{jk} = m_j \qquad (26)$$
$$R_m = (m_j - 1)^2 \qquad (27)$$
$$r_{jk} = \cos|\arccos(r_j) - 0| = r_j \qquad (28)$$
$$R_r = \frac{1-r_j^2}{r_j^2} m_j^2 \qquad (29)$$
$$R = (m_j - 1)^2 + \frac{1-r_j^2}{r_j^2} m_j^2 \qquad (30)$$
$$\gamma = \frac{1}{R} \qquad (31)$$

In other words, by considering Eq. (25) we study how a determistic relationship, which describes a noiseless channel, is transformed into the experimental relationship of a noisy channel, $m_j \neq 1$ and $|r_{jk}| < 1$. In conclusion, we can study a single scatterplot with the tools of the general theory. Therefore, the signal-to-noise ratio γ is still a single index that synthetically describes the relationship between x and y.

8.4. Performance of Linguistic Channels in Italian and in English

In this section we study the linguistic channels in Italian and in English. Table 6 reports the correlation and slope of the regression lines need to study the S–channel, I–channel, WI–channel and C–channel. Table 7 reports the signal-to-noise ratio in the channels.

It is very clear that the S–channel, I–channel and WI–channel between Italian (input) and English (output) are very noisy with very low signal-to-noise ratios. Only the S-Channel between Italian-E (input) and English (output) has a significant large $\Gamma \approx 27$ dB ($\gamma \approx 500$), even larger that the expected large $\Gamma \approx 24$ dB of the C–channel.

Table 6. Correlation and Slope of the regression line $y = mx$. Four digits are reported because some channels differ only at the third digit.

Text	S–Channel Sentences vs Words		I-Channel Words vs Interpunctions		WI–Channel Word Intervals vs Sentences		C–Channel Characters vs Words	
	Correlation Coefficient	Slope	Correlation Coefficient	Slope	Correlation Coefficient	Slope	Correlation Coefficient	Slope
Italian	0.5978	0.0470	0.9357	5.1220	0.7863	3.9374	0.9926	4.6263

Italian-E	0.6980	0.0649	0.9357	5.1220	0.8573	2.9070	0.9926	4.6263
English	0.7106	0.0623	0.9322	6.5785	0.8932	2.3587	0.9884	4.3560

Table 7. Signal-to-noise ratio Γ (dB) in the indicated channels.

	S-Channel			I-Channel			WI-Channel			C-Channel		
	Ita	Ita-M	Eng	Ita	Ita-M	Eng	Ita	Ita-M	Eng	Ita	Ita-M	Eng
Italian	∞	10.69	11.35	∞	∞	13.09	∞	8.11	2.50	∞	∞	23.08
Italian-E	7.48	∞	26.81	∞	∞	13.09	11.13	∞	12.04	∞	∞	23.08
English	8.36	27.22	∞	10.91	10.91	∞	7.56	14.06	∞	23.71	23.71	∞

To study the single scatterplots (i.e. the scatterplot between same variables) we need the correlation and slope of the regression lines reported in Table 4. Table 8 reports the signal-to-noise ratio in the channels.

It is very clear that only the channels related to characters and words – both charatcters and words are very much correlated, see Figure 6(b) and Table 6 – show large signal-to-noie ratios, while the other channels (with the exception of the sentences in Italian-E) are very noisy.

The conclusion regarding the linguistic channels is the same as above: a blind comparison would unlikely indicate the two texts refer to the same novel.

Table 8. Correlation and slope of the regression line $y = mx$, English versus Italian and Italian-E. Signal-to-noise ratio Γ (dB) and γ (linear) of the indicated linguistic channels.

Variable	Γ (dB)	γ
Characters	22.62	182.97
Paragraphs	12.62	18.27
Words	20.23	105.49
Sentences	6.45	4.42
Sentences (Italian-E)	15.65	36.75
Question Marks	11.27	13.40
Exclamation Marks	8.17	6.57
Commas	9.07	8.07
Interpunctions	12.35	17.19

In the next section, we consider a universal readability index. which measures the relative reading difficulty of texts.

9. Universal Readability Index

The universal readability index G_U – proposed in Ref. [64] - measures the relative reading difficulty of texts. It is underestood that readers are enough educated to understand the texts. It is given by:

$$G_U = 89 - 10kC_P + 300/P_F - 6(I_P - 6) \tag{32}$$

$$k = <C_{P,ITA}>/<C_{P,Lan}> \tag{33}$$

In the present paper $< C_{p,ITA} >= 4.62$ refers to Italian, $< C_{p,Lan} >= 4.36$ refers to English (see Table 5). By using Equations (22) and (33), the mean value $< kC_P >$ is forced to be equal to that found in Italian, namely 4.62. The rationale for this choice is that C_P is a parameter typical of a language/text which, if not scaled, would bias G_U without really quantifying the reading difficulty of readers, who in their own language are used, on average, to read shorter or longer words than in Italian. This scaling, therefore, avoids changing G_U only because a language has, on the average, words shorter (like English) than Italian. In any case, C_p affects Equation (32) much less than P_F or I_P.

Figure 14(a) shows the scatterplot of G_U between Italian and English (blue circles) and Italian-E and English (red circles). We can notice two interesting facts:

a) For $G_U < 55$, the readability of English is less than Italian; for $G_U \gtrsim 55$ the two values agree more, they align more with the 45° line (in Italian, in this range, the lower values of I_P balance in Eq. (32) the larger values of P_F).

b) If semicolons are replaced by periods, the Italian-E text would be much more readable than English, becasue P_F is noticeably reduced while I_P does not change.

These values could be "decoded, for example, into the minimum number of school years Y necessary to assess that a text/author is "easy" to read, according to the modern Italian school system, assumed as a common reference, Figure 14(b) [64].

This assumption does not mean, of course, that English readers should attend school for the same number of years of the Italian readers, but it is only a way to do relative comparisons, otherwise difficult to assess from the mere values of G_U. In other words, we should consider Y as an "equivalent" number of school years. We get, on the average, $Y \approx 9$ years for English, $Y \approx 8$ years for Italian and $Y \approx 7$ years for Italian-E.
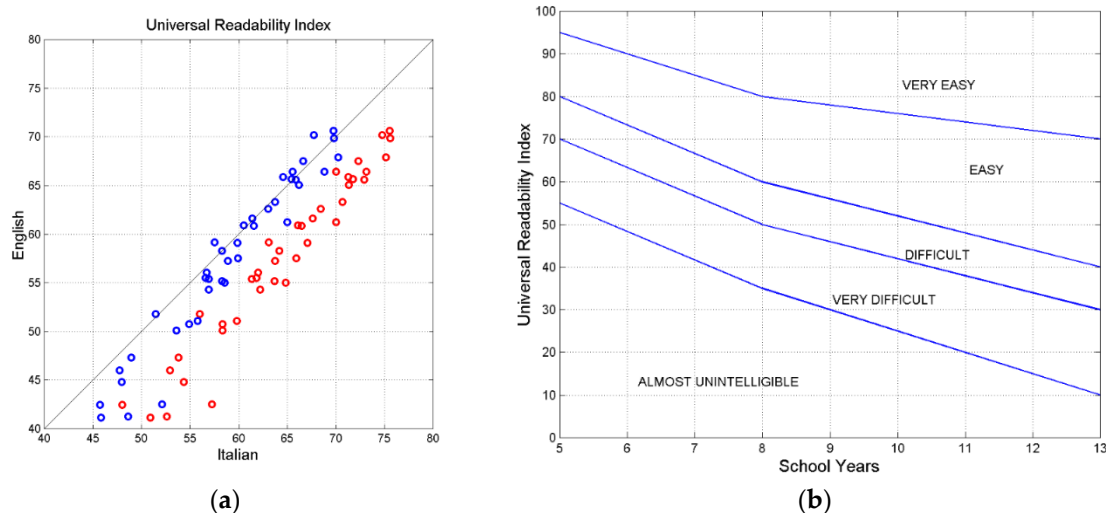
Figure 14. (a) Scatterplot between G_U in the two texts: blue indicates the comparison with Italian; red indicates the comparison with the Italian-E; (b) Chart to estimate the number of school years for a given G_U, according to text reading difficulty.

10. Discussion and Conclusions

We have mathematically compared the text of a famous Italian novel, *I promessi sposi* written by Alessandro Manzoni, to the text of a well-received masterful English translation, *The Betrothed* by Michael F. Moore. To avoid any misunderstanding, the mathematical theory and its multidimensional application to this translation does not measure or judge its efficacy and beauty, but only its mathematical underlying structure. In other words, the discussion and conclusion we reach are meant to be valid only in this context. We have measured the "equivalence" of the source

text and target text mathematically, just like we did with other unrelated couples of texts in any language, even different languages.

After the punctual discussion on the findings reported in each section, we can conclude that the translator largely modified the mathematical structure of the original Italian text. The translation theory he adopted was the "domestication" of the source text.

The almost total replacement of semicolons with periods was justified by stating that English is not as economical in its use of subject pronouns as Italian is. This was not the only change. Question and exclamation marks, and commas too, were also largely modified, with no clear justification due to an alleged "economy" of English. These modifications produce several consequences.

The short-term memory buffers required to the reader are significantly modified. The theoretical number of sentences allowed in Italian, i.e. the patterns of sentences available to Manzoni, is about two order of magnitude greater than in English, therefore the translator resused many times the same numerical pattern of words. Only when semicolons are replaced by periods in the Italian text, to simulate a reverse "domestication" translation, the two variables agree more.

The geometrical representation of texts and the related probability of error indicates that the two texts are practically disconnected and almost all linguistic channels are very noisy, with very poor signal-to-noise ratios, except the channel related to characters.

In conclusion, a blind comparison of the parameters of the multidemensional analysis on the two texts would unlikely indicate that the two texts refer to the same novel.

Appendix A. Incipit paragraph

Alessandro Manzoni, *I promessi sposi*, Capitolo 1.

«Quel ramo del lago di Como, che volge a mezzogiorno, tra due catene non interrotte di monti, tutto a seni e a golfi, a seconda dello sporgere e del rientrare di quelli, vien, quasi a un tratto, a ristringersi, e a prender corso e figura di fiume, tra un promontorio a destra, e un'ampia costiera dall'altra parte; e il ponte, che ivi congiunge le due rive, par che renda ancor più sensibile all'occhio questa trasformazione, e segni il punto in cui il lago cessa, e l'Adda rincomincia, per ripigliar poi nome di lago dove le rive, allontanandosi di nuovo, lascian l'acqua distendersi e rallentarsi in nuovi golfi e in nuovi seni. La costiera, formata dal deposito di tre grossi torrenti, scende appoggiata a due monti contigui, l'uno detto di san Martino, l'altro, con voce lombarda, il Resegone, dai molti suoi cocuzzoli in fila, che in vero lo fanno somigliare a una sega: talché non è chi, al primo vederlo, purché sia di fronte, come per esempio di su le mura di Milano che guardano a settentrione, non lo discerna tosto, a un tal contrassegno, in quella lunga e vasta giogaia, dagli altri monti di nome più oscuro e di forma più comune. Per un buon pezzo, la costa sale con un pendìo lento e continuo; poi si rompe in poggi e in valloncelli, in erte e in ispianate, secondo l'ossatura de' due monti, e il lavoro dell'acque. Il lembo estremo, tagliato dalle foci de' torrenti, è quasi tutto ghiaia e ciottoloni; il resto, campi e vigne, sparse di terre, di ville, di casali; in qualche parte boschi, che si prolungano su per la montagna. Lecco, la principale di quelle terre, e che dà nome al territorio, giace poco discosto dal ponte, alla riva del lago, anzi viene in parte a trovarsi nel lago stesso, quando questo ingrossa: un gran borgo al giorno d'oggi, e che s'incammina a diventar città. Ai tempi in cui accaddero i fatti che prendiamo a raccontare, quel borgo, già considerabile, era anche un castello, e aveva perciò l'onore d'alloggiare un comandante, e il vantaggio di possedere una stabile guarnigione di soldati spagnoli, che insegnavan la modestia alle fanciulle e alle donne del paese, accarezzavan di tempo in tempo le spalle a qualche marito, a qualche padre; e, sul finir dell'estate, non mancavan mai di spandersi nelle vigne, per diradar l'uve, e alleggerire a' contadini le fatiche della vendemmia. Dall'una all'altra di quelle terre,

dall'alture alla riva, da un poggio all'altro, correvano, e corrono tuttavia, strade e stradette, più o men ripide, o piane; ogni tanto affondate, sepolte tra due muri, donde, alzando lo sguardo, non iscoprite che un pezzo di cielo e qualche vetta di monte; ogni tanto elevate su terrapieni aperti: e da qui la vista spazia per prospetti più o meno estesi, ma ricchi sempre e sempre qualcosa nuovi, secondo che i diversi punti piglian più o meno della vasta scena circostante, e secondo che questa o quella parte campeggia o si scorcia, spunta o sparisce a vicenda. Dove un pezzo, dove un altro, dove una lunga distesa di quel vasto e variato specchio dell'acqua; di qua lago, chiuso all'estremità o piuttosto smarrito in un gruppo, in un andirivieni di montagne, e di mano in mano più allargato tra altri monti che si spiegano, a uno a uno, allo sguardo, e che l'acqua riflette capovolti, co' paesetti posti sulle rive; di là braccio di fiume, poi lago, poi fiume ancora, che va a perdersi in lucido serpeggiamento pur tra' monti che l'accompagnano, degradando via via, e perdendosi quasi anch'essi nell'orizzonte. Il luogo stesso da dove contemplate que' vari spettacoli, vi fa spettacolo da ogni parte: il monte di cui passeggiate le falde, vi svolge, al di sopra, d'intorno, le sue cime e le balze, distinte, rilevate, mutabili quasi a ogni passo, aprendosi e contornandosi in gioghi ciò che v'era sembrato prima un sol giogo, e comparendo in vetta ciò che poco innanzi vi si rappresentava sulla costa: e l'ameno, il domestico di quelle falde tempera gradevolmente il selvaggio, e orna vie più il magnifico dell'altre vedute.

Per una di queste stradicciole, tornava bel bello dalla passeggiata verso casa, sulla sera del giorno 7 novembre dell'anno 1628, don Abbondio…»

Alessandro Manzoni, *The Betrothed*, Chapter 1. Translated by Moore

«The branch of Lake Como that turns south between two unbroken mountain chains, bordered by coves and inlets that echo the furrowed slopes, suddenly narrows to take the flow and shape of a river, between a promontory on the right and a wide shoreline on the opposite side. The bridge that joins the two sides at this point seems to make this transformation even more visible to the eye and mark the spot where the lake ends and the Adda begins again, to reclaim the name lake where the shores, newly distant, allow the water to spread and slowly pool into fresh inlets and coves. Formed from the sediment of three large streams, the shoreline lies at the foot of two neighboring mountains, the first called San Martino, the second, in Lombard dialect, the "Resegone"—the big saw—after the row of many small peaks that really do make it look like one. So clear is the resemblance that no one—provided they are directly facing it, from the northern walls of Milan, for example—could fail to immediately distinguish this summit from other mountains in that long and vast range with more obscure names and more common shapes. For a good stretch the shore rises upward into a smooth rolling slope. Then it breaks off into small hills and ravines, steep inclines and flat terraces, molded by the contours of the two mountains and the erosion of the waters. The water's edge, cut by the outlets of the streams, is almost all pebbles and stones. The rest is fields and vineyards, dotted with towns, villages, and hamlets. Here and there a woods climbs up the side of the mountain.

Lecco, the capital that lends its name to the province, stands at a short remove from the bridge, and is on and indeed partly inside the lake when the water rises. Nowadays it is a large town well on its way to becoming a city. At the time of the events I am about to relate, this already good-sized village was also fortified, which conferred upon it the honor of a commander in residence, and the benefit of a permanent garrison of Spanish soldiers, who taught modesty to the girls and women of the town, gave an occasional tap on the back to a husband or father, and, at summer's end, never failed to spread out into the vineyards to thin the grapes and relieve the peasants of the trouble of harvesting them.

Roads and footpaths used to run—and still do—from town to town, from summit to shore, from hill to hill. Some are more or less steep; some are level. In places they dip, sinking between two walls, and all you can see when you look up is a patch of sky and a mountaintop. In others they climb to open embankments, where the view encompasses a broader panorama, always rich, always new, depending on the vantage point and how much of the vast expanse can be seen, and on whether the landscape protrudes or recedes, stands out from or disappears into the horizon.

One piece, then another, then a long stretch of that vast and varied mirror of water. Over here, the lake—terminating at the far end or vanishing into a cluster, a procession of mountains—slowly

growing wider between yet more mountains that unfold before the eyes, one by one, whose image, alongside that of the towns by the lake, is reflected upside down in the water. Over there, a bend in the river, then more lake, then river again disappearing into a shiny ribbon curling between the mountains on either side, and slowly descending to vanish on the horizon.

The place from which you contemplate these varied sights offers its own display on every side. The mountain along whose slopes you walk unfolds its peaks and crags above and around you, distinct, prominent, changing with every step, opening and then circling into ridges where a lone summit had at first appeared. The shapes that were reflected in the lake only minutes before now appear close to the summit. The tame, pleasant foothills temper the wild landscape and enhance the magnificence of the other vistas.

Along one of these footpaths, on the evening of the seventh day of November in 1628, Don Abbondio…»

Appendix B. Example of Texts

Example in which the number of question marks do not coincide, see the text in italics.

Alessandro Manzoni, *I promessi sposi*, Capitolo 21.

«Oh Signore! *pretendere!* Cosa posso pretendere io meschina, se non che lei mi usi misericordia? Dio perdona tante cose, per un'opera di misericordia! Mi lasci andare; per carità mi lasci andare! *Non torna conto a uno che un giorno deve morire di far patir tanto una povera creatura.* Oh! lei che può comandare, dica che mi lascino andare! M'hanno portata qui per forza. Mi mandi con questa donna a *** dov'è mia madre. Oh Vergine santissima! mia madre! mia madre, per carità, mia madre! Forse non è lontana di qui... ho veduto i miei monti! Perché lei mi fa patire? Mi faccia condurre in una chiesa. Pregherò per lei, tutta la mia vita. Cosa le costa dire una parola? Oh ecco! vedo che si move a compassione: dica una parola, la dica. Dio perdona tante cose, per un'opera di misericordia!»

Alessandro Manzoni, *The Betrothed*, Chapter 21.

«Oh, sir! *What do I expect?* What can a poor girl like me expect? Only your mercy. God will forgive many things for an act of mercy! Let me go, for the love of God, let me go! *You, who will also have to face your maker one day: What do you have to gain from causing a poor creature so much suffering?* You have the power to give orders: So tell them to let me go! They brought me here by force. Send me with this woman to *** , where my mother lives. Oh Blessed Virgin! My mother! My mother, for goodness' sake, my mother! Perhaps it is not far from here. I saw my mountains! Why are you making me suffer? Tell them to take me to a church. I will pray for you for the rest of my life. What would it cost you to say one word? Ah, yes! I can see that you are moved to pity: Say the word, say it. God will forgive many things for an act of mercy!»

Appendix C. List of Mathematical Symbols and Meaning

Symbol	Definition
C_P	Characters per word
G_U	Universal readability index
I_M	Mismatch index
I_p	Word interval
M_F	Word intervals per sentence
P_F	Words per sentence
R	Noise–to–signal ratio
R_m	Regression noise–to–signal ratio
R_r	Correlation noise–to–signal ratio
S	Total number of sentences

W	Total number of words
n_C	Number of characters
n_W	Number of words
n_S	Number of sentences
n_I	Number of interpunctions
n_{I_P}	Number of word intervals
γ	Signal–to–noise ratio
Γ	Signal–to–noise ratio (dB)
m_{jk}	Slope of regression line of text j versus text k
r_{jk}	Correlation coefficient between text j and text k

Appendix D. Inequalities

Let M be the number of samples (i.e., number of disjoint blocks of text, such as chapters or books), then, for example, the statistical mean value $< P_F >$, is given by

$$< P_F >= \sum_{k=1}^{M} P_{F,k} \times \left(n_{W,k}/W \right) \tag{A1}$$

where $W = \sum_{k=1}^{M} n_{W,k}$ is the total number of words. Notice that $< P_F > \neq \frac{1}{M}\sum_{k=1}^{M} P_{F,k} \neq \sum_{k=1}^{M} n_{W,k}/\sum_{k=1}^{M} n_{S,k}=W/S$, where S is the total number of sentences.

In general, the average values calculated from sample totals are always smaller than their statistical means, therefore they give lower bounds, as I prove in the following.

Let us consider, for example, the parameter P_F. Because of Chebyshev's inequality ([65], inequality 3.2.7), we can write Eq.(A1) as:

$$< P_F >= \sum_{k=1}^{M} \frac{n_{W_k}}{n_{S_k}} \frac{n_{W_k}}{W} \geq \frac{1}{M}\sum_{k=1}^{M} \frac{n_{W_k}}{n_{S_k}} \sum_{k=1}^{M} \frac{n_{W_k}}{W} = \frac{1}{M}\sum_{k=1}^{M} \frac{n_{W_k}}{n_{S_k}} \tag{A2}$$

Eq.(A2) states that the mean calculated with samples weighted $1/M$ (arithmetic mean) is smaller than (or equal to) the mean calculated with samples weighted n_{W_k}/W.

Now, again for Chebyshev's inequality, we get:

$$\sum_{k=1}^{M} \frac{n_{W_k}}{n_{S_k}} \geq \frac{1}{M}\sum_{k=1}^{M} n_{W_k} \sum_{k=1}^{M} \frac{1}{n_{S_k}} = \frac{W}{M}\sum_{k=1}^{M} \frac{1}{n_{S_k}} \tag{A3}$$

Further, for Cauchy–Schawarz's inequality (or by the fact that the harmonic mean is less than, or equal to, the arithmetic mean), we get:

$$\sum_{k=1}^{M} \frac{1}{n_{S_k}} \geq \frac{M^2}{\sum_{k=1}^{M} n_{S_k}} \tag{A4}$$

Finally, by inserting these inequalities in (6), we get:

$$< P_F > \geq \frac{W}{M^2} \frac{M^2}{\sum_{k=1}^{M} n_{S_k}} = \frac{W}{S} \tag{A5}$$

Eq.(A5) establishes that the statistical mean calculated with samples weighted n_{W_k}/W is greater than (or equal to) the average calculated with sample totals. The values given by these three methods of calculation coincide only if all texts are perfectly identical, i.e. with the same number of characters, words, sentences and interpunctions, a case improbable.

References

1. Matricciani, E. Deep Language Statistics of Italian throughout Seven Centuries of Literature and Empirical Connections with Miller's 7 ∓ 2 Law and Short–Term Memory. *Open J. Stat.* **2019**, *9*, 373–406. https://doi.org/10.4236/ojs.2019.93026.

2. Matricciani, E. A Statistical Theory of Language Translation Based on Communication Theory. *Open J. Stat.* **2020**, *10*, 936–997. https://doi.org/10.4236/ojs.2020.106055.

3. Matricciani, E. Multiple Communication Channels in Literary Texts. *Open J. Stat.* **2022**, *12*, 486–520. https://doi.org/10.4236/ojs.2022.124030.

4. Matricciani, E. Linguistic Mathematical Relationships Saved or Lost in Translating Texts: Extension of the Statistical Theory of Translation and Its Application to the New Testament. *Information* **2022**, *13*, 20. https://doi.org/10.3390/info13010020.

5. Matricciani, E. Capacity of Linguistic Communication Channels in Literary Texts: Application to Charles Dickens' Novels. *Information* **2023**, *14*, 68. https://doi.org/10.3390/info14020068.

6. Matricciani, E. Linguistic Communication Channels Reveal Connections between Texts: The New Testament and Greek Literature. *Information* **2023**, *14*, 405. https://doi.org/10.3390/info14070405.

7. Matricciani, E. Is Short–Term Memory Made of Two Processing Units? Clues from Italian and English Literatures down Several Centuries. *Information* **2024**, *15*, 6. https://doi.org/10.3390/info15010006.

8. Matricciani, E. A Mathematical Structure Underlying Sentences and Its Connection with Short–Term Memory. *Appl. Math* **2024**, *4*, 120–142. https://doi.org/10.3390/appliedmath4010007.

9. Catford J.C.; *A linguistic theory of translation. An Essay in Applied Linguistics*. **1965**, Oxford Univeristy Press.

10. Munday, J.; *Introducing Translation studies. Theories and applications*, **2008**, Routledge, New York.

11. Proshina, Z., *Theory of Translation*, **2008**, Far Eastern University Press.

12. Warren, R. (ed.), *The Art of Translation: Voices from the Field*, **1989**, Boston, MA: North-eastern University Press.

13. Wilss, W., *Knowledge and Skills in Translator Behaviour*, **1996**, Amsterdam and Philadelphia: John Benjamins.

14. Shannon, C.E. A Mathematical Theory of Communication, *The Bell System Technical Journal*, **1948**, 27, p.379–423, and p. 623–656.

15. Hyde, G.M., Literary Translation, *Hungarian Studies in English*, **1991**, 22 , 39-47.

16. Yousef, T. Literary Translation: Old and New Challenges, *International Journal of Arabic-Eng lish Studies* (IJAES), **2012**, 13, 49-64.

17. Nũñez, K.J., Literary translation as an act of mediation between author and reader, *Estudios de Traducción*, **2012**, 2, 21-31.

18. Bernaerts, L., De Bleeker, L., De Wilde, J., Narration and translation, *Language and Literature*, **2014**, 23(3) 203–212 DOI: 10.1177/0963947014536504.

19. Ghazala, H.S., Literary Translation from a Stylistic Perspective, *Studies in English Language Teaching*, **2015**, 3, 2, 124-145.

20. Suo, X., A New Perspective on Literary Translation. Strategies Based on Skopos Theory, *Theory and Practice in Language Studies*, **2015**, 5, 1, 176-183, http://dx.doi.org/10.17507/tpls.0501.24

21. Munday, J. (Editor), *The Routledge Companion to Translation Studies*, **2009**, Routledge, New York.

22. Munday, J., *Introducing Translation Studies.Theories and applications*, **2016**, Routledge, New York.

23. Panou, D. Equivalence in Translation Theories: A Critical Evaluation, *Theory and Practice in Language Studies*, **2013**, 3, 1, 1-6.

24. Saule, B., Aisulu, N. Problems of translation theory and practice: original and translated text equivalence, *Procedia - Social and Behavioral Sciences*, **2014,**136, 119-123.

25. Krein- Kühle, M., Translation and Equivalence, in House, J (Ed.), *Translation: A Multidisciplinary Approach*, **2014**, Palgrave Advances in Language and Linguistics.

26. Matricciani, E. Multi–Dimensional Data Analysis of Deep Language in J.R.R. Tolkien and C.S. Lewis Reveals Tight Mathematical Connections. *AppliedMath* **2024**, *4*, 927–949. https://doi.org/10.3390/appliedmath4030050.

27. Venuti, L., *The Translator's Invisibility. A History of Translation*, **1995**, Routledge, New York.

28. Manzoni, A. *The Betrothed*, translated and with an Introduction by Michael F. Moore, 2022, New York: The Modern Library.

29. Cella, R., *Storia dell'italiano*, Bologna: Società Editrice Il Mulino, **2015**.

30. Frare, P., Una struttura in movimento: sulla forma artistica dei «I promessi sposi». *The Italianist*, **1996**, 16(1), 62–75. https://doi.org/10.1179/ita.1996.16.1.62.

31. Raboni, G., La scrittura purgata: sulla cronologia della seconda minuta dei Promessi sposi, *Filologia italiana*, **2008**, 5, Istituti editoriali e poligrafici internazionali;

32. Frare, P., *La scrittura dell'inquietudine: saggio su Alessandro Manzoni*, **2006**, Firenze : L.S. Olschki.

33. *Frare, P., Salvioli, M., Prodigi di misericordia e forza del linguaggio. Sui capitoli XXI e XXIII dei «Promessi sposi», Munera, 2016; 3, 109-119.*

34. Bruthiaux, P., The Rise and Fall of the Semicolon: English Punctuation Theory and English Teaching Practice, *Applied Linguistics,* **1995**, 16, 1, 1-14.

35. Duncan, M., Whatever Happened to the Paragraph? *College English*, **2007**, 69, 5, 470-495.

36. Watson, C., Points of Contention: Rethinking the Past, Present, and Future of Punctuation, *Critical Inquiry*, **2012**, 38, 3, 649-672.

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

38. Papoulis, A.; *Probability & Statistics*, **1990**, Prentice Hall, 1990.

39. Lindgren, B.W. *Statistical Theory*, 2nd ed.; **1968**, MacMillan Company: New York, NY, USA.

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

41. Crowder, R.G. Short–term memory: Where do we stand?, **1993**, *Memory & Cognition*, 21, 142–145.

42. Lisman, J.E., Idiart, M.A.P. Storage of 7 ± 2 Short–Term Memories in Oscillatory Subcycles, **1995,** *Science*, 267, 5203, 1512– 1515.

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

44. Bachelder, B.L. The Magical Number 7 ± 2: Span Theory on Capacity Limitations. *Behavioral and Brain Sciences* **2001**, 24, 116–117. https://doi.org/10.1017/S0140525X01243921

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

46. Burgess, N., Hitch, G.J. A revised model of short–term memory and long–term learning of verbal sequences, **2006**, *Journal of Memory and Language*, 55, 627–652.

47. Richardson, J.T.E, Measures of short–term memory: A historical review, **2007**, *Cortex*, 43, 5, 635–650.

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

49. Melton, A.W., Implications of Short–Term Memory for a General Theory of Memory, **1963**, *Journal of Verbal Learning and Verbal Behavior*, 2, 1–21.

50. Atkinson, R.C., Shiffrin, R.M., The Control of Short–Term Memory, **1971**, *Scientific American*, 225, 2, 82–91.

51. Murdock, B.B. Short–Term Memory, **1972**, *Psychology of Learning and Motivation*, 5, 67–127.

52. Baddeley, A.D., Thomson, N., Buchanan, M., Word Length and the Structure of Short–Term Memory, *Journal of Verbal Learning and Verbal Behavior*, **1975**, 14, 575–589.

53. Case, R., Midian Kurland, D., Goldberg, J. Operational efficiency and the growth of short–term memory span, 1982, Journal of Experimental Child Psychology**,** 33, 386–404.

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

55. Pothos, E.M., Joula, P., Linguistic structure and short–term memory, *Behavioral and Brain Sciences*, **2000**, 124, 138–139.

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

57. Jonides, J., Lewis, R.L., Nee, D.E., Lustig, C.A., Berman, M.G., Moore, K.S., The Mind and Brain of Short–Term Memory, **2008** *Annual Review of Psychology*, 69, 193–224.

58. Potter, M.C. Conceptual short–term memory in perception and thought, **2012**, *Frontiers in Psychology*, doi: 10.3389/fpsyg.2012.00113.

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

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

61. Norris, D., Short–Term Memory and Long–Term Memory Are Still Different, **2017**, *Psychological Bulletin*, 143, 9, 992–1009.

62. Houdt, G.V., Mosquera, C., Napoles, G., A review on the long short–term memory model, **2020**, *Artificial Intelligence Review*, 53, 5929–5955.

63. Islam, M., Sarkar, A., Hossain, M., Ahmed, M., Ferdous, A. Prediction of Attention and Short–Term Memory Loss by EEG Workload Estimation. *Journal of Biosciences and Medicines*, **2023**, 304–318. doi: 10.4236/jbm.2023.114022.

64. Matricciani, E. Readability across Time and Languages: The Case of Matthew's Gospel Translations. *AppliedMath* **2023**, *3*, 497–509. https://doi.org/10.3390/appliedmath3020026.

65. Abramovitz, M.; Stegun, I.A., *Handbook of Mathematical Formulas*, Dover publications, **1972**, New York, 9th ed.