ICJK

Journal of Mathematical Studies of Literature
http://dx.doi.org/10.62762/JMSL.2026.487587

RESEARCH ARTICLE

A Multidimensional Mathematical Analysis of Dante's Divina Commedia

Emilio Matricciani[1,*]

[1] Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano, 20133 Milan, Italy

Abstract

Dante Alighieri's Divina Commedia is the preeminent work of the Italian literature and one of the greatest of the world literature. The poem allegorically represents the soul's journey toward God passing through *Inferno*, *Purgatorio* and *Paradiso*. At each stage of the journey, Dante adapts his language to the different contexts he describes, bringing together, in a single great work, the various types of language that had previously been specific to comedy or tragedy. This is why scholars speak of pluristylism. Compared to *Inferno* and *Purgatorio*, the language of *Paradiso* increases in complexity and in linguistic perfection to express profound theological concepts, making the reading more demanding and difficult to interpret. Since scholars unanimously find notable differences between *Paradiso* on the one hand, and *Inferno* and *Purgatorio* on the other, in this article I have examined whether the mathematical structure of deep language, of which Dante is unaware, is different in *Inferno*, *Purgatorio* and *Paradiso*. The answer is affirmative. The multidimensional mathematical analysis – based on deep–language variables and short–term memory equivalent modelling – shows that *Inferno* and *Purgatorio* are very similar and markedly different from *Paradiso*. The Divina Commedia was written in hendecasyllabic verse by counting syllables. Each verse, made on the average by 7 words, matches the central value of Miller's 7±2 Law. The approach of scholars using traditional tools, and the approach of scientists using mathematical tools can reinforce each other in the more objective evaluation of literary texts, especially when comparing texts written by different authors or by the same author.

Keywords: alphabetic languages, short–term memory, geometric representation, linguistic variables, likeness index, readability index, *Inferno*, *Purgatorio*, *Paradiso*, Dante's tercets, hendecasyllabic verses, Petrarca.

1 Introduction: Dante's Divina Commedia: a masterpiece of the world literature

Dante Alighieri's Divina Commedia (Divine Comedy, written between 1304 and 1321) is the preeminent work of the Italian literature and a masterpiece of the world literature. Dante Alighieri (1265–1321), with Giovanni Boccaccio (1313–1375) and Francesco Petrarca (1304–1374), established the Tuscan dialect as the Italian language [1–10].

The Divina Commedia explores the condition of the

Submitted: 09 April 2026
Accepted: 08 June 2026
Published: 18 June 2026

Vol. 1, **No.** 1, 2026.
10.62762/JMSL.2026.487587

***Corresponding author:**
✉ Emilio Matricciani
Emilio.Matricciani@polimi.it

Citation
Matricciani, E. (2026). A Multidimensional Mathematical Analysis of Dante's Divina Commedia. *Journal of Mathematical Studies of Literature*, 1(1), 4–20.

soul after death and portrays a vision of divine justice, in which individuals receive punishment or reward based on their actions. It describes Dante's journey through *Inferno* (Hell), *Purgatorio* (Purgatory), and *Paradiso* (Paradise). The poem allegorically represents the soul's journey toward God, beginning with the recognition and rejection of sin (*Inferno*), followed by the Christian life of penance (*Purgatorio*), and finally the soul's ascension to God (*Paradiso*). I indicate with the Italian terms both the three texts (cantiche) in which the Divina Commedia is divided and the three heavenly places to which they correspond.

On his journey, three characters guide Dante: Virgil, Beatrice and Bernard of Clairvaux. Virgil – representing human reason – leads him throughout *Inferno* and much of *Purgatorio*. Beatrice – representing divine revelation, theology, grace, and faith – leads him from the end of *Purgatorio* to the beginning of *Paradiso*; Bernard of Clairvaux – representing contemplative mysticism and devotion to Mary, Mother of Jesus – leads him to the final part of *Paradiso*.

Inferno is a place of sin and suffering, *Purgatorio* is a place of redemption, and finally *Paradiso* is the abode of the blessed. At each stage of the journey, he encounters different realities, from the lowest, in *Inferno*, to the highest, in *Paradiso*.

For the purposes of this article, Dante adapts his language to the different contexts he wishes to describe. Depending on the characters he converses with, Dante's vocabulary ranges from everyday language to scientific, from the popular to the refined, used by the most cultured people. The linguistic expressions that enrich his verses also vary. As Dante's language evolves, as his journey towards *Paradiso* progresses, so does his narrative register. For this reason, scholars of the Italian Literature [1–10] describe the style of the Divina Commedia as pluristylism. Dante achieved a literary revolution, bringing together in a single great work the various types of language that had previously been specific to comedy or tragedy.

Scholars of Italian literature are unanimous in their remarks on Dante's language in the three cantiche. In *Paradiso*, the language is extremely lofty because the subject matter is extremely difficult. At the beginning of *Paradiso*, Dante states that he needs special help, a powerful poetic inspiration. He adds that his readers also need particularly serious cultural tools. The lofty and difficult language of *Paradiso* is not an end in itself but is caused by the lofty and previously untackled subject matter.

The scholars speak of pluristylism because Dante adapts his language to the different scenarios he describes. In *Paradiso*, he even reaches the point of expressing himself with sublime language, to help the reader understand what is difficult, almost impossible to convey in words: the transcendental vision of God.

As Dante's language gradually elevates as the journey toward *Paradiso* progresses, so does the narrative register. This is why scholars speak of pluristylism referring to the Divina Commedia.

Compared to *Inferno* and *Purgatorio*, the language of *Paradiso* clearly increases in complexity, with a quest for absolute language perfection to express profoundly significant theological concepts. The reading thus becomes more demanding and difficult to interpret, rich in metaphors, with a reduction in encounters with souls to leave space for the moral and religious explanations of Beatrice, Dante's beloved woman and emblem of Theology.

Since scholars do signal a large difference between *Paradiso*, on one hand, and *Inferno* and *Purgatorio* on the other hand, on many levels, the purpose of this article is to examine whether the deep–language mathematical structure, of which Dante is unaware, is different in *Inferno*, *Purgatorio*, and *Paradiso*. The answer is affirmative. The mathematical structure of *Paradiso* is clearly different from that of *Inferno* and *Purgatorio*, as I show next.

I developed the mathematical tools used in analyzing the Divina Commedia in the past seven years, building on Miller's classic capacity-limit framework [27], and have applied them in several study–cases [11–17]. These tools are very different from those used by other scholars pursuing quantitative investigations of the poem. Some of these scholars are interested in word frequency and other lexical markers, using multivariate statistical classification or grapheme-level Markov modelling to separate the three cantiche [18, 20]; others focus instead on the numerical and symbolic organisation of the text, such as canto-length numerology or the recurrence of structurally significant numbers [19, 21]. None of them has explored the mathematical structure underlying the texts, of which the authors are unaware and which, therefore, reveals objective relationships. On the contrary, the mathematical tools I have developed do that.

After this introductory section, Section 2 describes the poetic structure of the Divina Commedia as Dante

designed it. Section 3 defines the deep–language variables used in this article. Section 4 models the probability density function of deep–language variables and calculates the likeness index, which measures the probability that a text of *Inferno* might be confused, mathematically, with *Purgatorio* or *Paradiso*, or vice–versa. Section 5 recalls and applies the universal readability index to the three Cantiche to show which one is more readable. Section 6 recalls and apply a geometric representation of alphabetic texts useful to synthesize graphically their mathematical "imprint" and compare them. Section 7 recalls the modelling of the extended short–term memory with three equivalent buffers and applies it to the Divina Commedia. Section 8 summarizes the main results and indicates future work. Appendix A describes how the textual data have been obtained and processed and Appendix B reports the list of mathematical symbols used in the article.

2 The poetic structure of the Divina Commedia

The fundamental textual structure of the Divina Commedia is the *canto* (plural: *canti*). The poem is composed of 100 canti, grouped into three sections, or *cantiche* (plural of *cantica*), namely *Inferno*, *Purgatorio*, and *Paradiso*. Each cantica contains 33 canti, plus an additional canto at the beginning of *Inferno*, which serves as an introduction to the entire poem. The Divina Commedia is based on the Christian symbolism of the number 3 (Father, Son and Holy Spirit) and its multiples, of 1 (the one God) and 100 (the totality of God) [19].

The canti are written in chained lines (*stanze*, plural of *stanza*) of hendecasyllabic verses, subdivided in tercets (*terzine*, plural of *terzina*).

The hendecasyllable is undoubtedly the most important building block of Italian metrical poetry. The greatest Italian poets, including the fathers of the Italian language, as Dante, have extensively used this metrical structure. The name hendecasyllable might perhaps mislead one into thinking that it is always characterized by being made up of 11 syllables, but this is not always the case because its structure can take various forms, with syllables ranging from 10 to 12. Its defining feature is a constant stress on the tenth syllable, so that the number of syllables in the verse may vary.

The terzina (tercet) is an Italian metrical form composed of three stanze, or lines. In the complex metrical pattern of the tercets, the first and third lines of the first tercet rhyme. Furthermore, the second line rhymes with the first and third lines of the following stanza. Finally, the entire canto ends with a line that rhymes with the second line of the last complete stanza. Dante was the first poet to use the tercet for a long poem, therefore, it is known as Dante's tercet.

The structure of the tercets is shown in Table 1, which reports the beginning of the introductory canto of *Inferno*, the most celebrated incipit in the Italian Literature that generations of Italian High School students have memorized and still do.

In Appendix A, I describe how the raw linguistic data were calculated in the best available digital textual source. These data are the basis of the elaborations reported below.

Table 1. Introductory canto of the *Divina Commedia* (*Inferno*), the most famous incipit in Italian Literature.

Line	Verse	Rhyme
1	Nel mezzo del cammin di nostra vita	A
2	mi ritrovai per una selva oscura	B
3	ché la diritta via era smarrita.	A
4	Ahi quanto a dir qual era è cosa dura	B
5	esta selva selvaggia e aspra e forte	C
6	che nel pensier rinova la paura!	B
7	Tant'è amara che poco è più morte;	C
8	ma per trattar del ben ch'i' vi trovai,	D
9	dirò de l'altre cose ch'i' v'ho scorte.	C
10	Io non so ben ridir com'i' v'intrai,	D
11	tant'era pien di sonno a quel punto	E
12	che la verace via abbandonai.	D

Each canto is written in chained hendecasyllabic verses (*stanze*), called tercet, an Italian metrical form composed of three-line *stanze*. In the complex metrical pattern of the tercet, the first and third lines of the first tercet rhyme. Furthermore, the second line rhymes with the first and third lines of the following stanza, and so on. The colored part of the hendecasyllabic verses highlights the rhyme structure.

The combination of the tercets and hendecasyllabic verses present in the stanze makes the distribution of the words per cantica very similar, as Table 2 shows. On the contrary, the number of sentences per cantica is different, decreasing from *Inferno* to *Paradiso*.

The tercet structure of the hendecasyllabic verses leads to more compact statistics concerning syllables, characters and words, as shown in Table 3, which

Table 2. Mean and standard deviation of the words and sentences contained in each *cantica*.

Cantica	Words	Sentences
Inferno	974.2 ± 68.3	41.3 ± 6.5
Purgatorio	1001.7 ± 49.6	39.8 ± 4.8
Paradiso	975.0 ± 39.5	32.2 ± 4.2

evidences that the words in *Paradiso* are on average longer than in *Inferno* and *Purgatorio*. More on this topic in the next section.

Table 3. Mean and standard deviation of syllables and characters in the three *cantiche*.

Cantica	Syllables per word	Characters per syllable	Syllables per line
Inferno	1.69 ± 0.04	2.45 ± 0.03	11.85 ± 0.10
Purgatorio	1.70 ± 0.04	2.44 ± 0.02	11.87 ± 0.10
Paradiso	1.74 ± 0.03	2.43 ± 0.02	11.77 ± 0.08

3 Deep–language variables

In this section, I recall some fundamental deep-language variables [11]. These variables are very useful for comparing texts because writers are mostly unaware of them, therefore they can assess objectively the similarity of texts beyond writers' awareness. These variables refer to the "surface" structure of texts, not to the "deep" structure mentioned in cognitive theory.

Let n_C, n_W, n_S and n_I be respectively the number of characters, words, sentences and interpunctions (punctuation marks) calculated in disjoint blocks of texts, the canti in this article. We can define four deep-language variables (Appendix B lists the mathematical symbols used in the present paper), namely:

Number of characters per word, C_P:

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

The number of words per sentence, P_F:

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

Number of words per interruption, referred to as the word interval, I_P:

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

Number of word intervals per sentence, M_F:

$$M_F = \frac{n_I}{n_S} \tag{4}$$

Notice that I_P, P_F and M_F, are not independent because:

$$I_P = P_F / M_F \tag{5}$$

In some of my previous papers, because of unfortunate typos, I_P was wrongly written as $I_P = \frac{n_I}{n_W}$, i.e. the reciprocal of Eq.(3) instead of Eq. (3). These typos, however, had no effect either on the processing of the textual data or on the findings described in these papers because the correct definition was always used in the data processing.

We can define, for poetry, the number of words per line, $I_{P,\ell}$, to which I refer as the word interval per line. Defined the number of words per canto, w_{can}, and the number of lines in the canto, ℓ_{can}, $I_{P,\ell}$ is defined by:

$$I_{P,\ell} = \frac{w_{can}}{\ell_{can}} \tag{6}$$

Table 4 reports the mean and standard deviation of these variables. These statistics are calculated from 34 samples in *Inferno*, 33 samples both in *Purgatorio* and in *Paradiso*.

These results are interesting. We can notice, passing from *Inferno* to *Purgatorio* and to *Paradiso*, that all deep-language variables show an increasing trend. In particular, considering the mean value:

- C_P: Words become slightly longer; in *Paradiso* 3% longer than in *Inferno*.
- P_F: Sentences in *Paradiso* are 27% longer than in *Inferno*, and 20% longer than in *Purgatorio*.
- I_P: The word interval is the same in *Inferno* and in *Purgatorio*, but 13% longer in *Paradiso*.
- M_F: The number of word intervals per sentence in *Paradiso* is 13% larger than in *Inferno* and 6% larger than in *Purgatorio*.
- $I_{P,\ell}$: There is no significant difference in the three cantiche; $I_{P,\ell} \approx 7$ is a very interesting result because this value is the center value of Miller's Law (Section 7).

In conclusion, the three cantiche are mathematically diverse, with the largest difference found in *Paradiso*. A clear increasing trend is present from *Inferno* to *Paradiso*. These differences have consequences on the

Table 4. Mean value and standard deviation of the indicated deep-language variables in the three cantiche.

Cantica	$< \mathbf{C_P} > \pm \mathbf{s}$	$< \mathbf{P_F} > \pm \mathbf{s}$	$< \mathbf{I_P} > \pm \mathbf{s}$	$< \mathbf{M_F} > \pm \mathbf{s}$	$< \mathbf{I_{P,\ell}} > \pm \mathbf{s}$
Inferno	4.13 ± 0.07	24.08 ± 3.65	6.38 ± 0.48	3.77 ± 0.44	7.02 ± 0.16
Purgatorio	4.16 ± 0.10	25.51 ± 2.99	6.37 ± 0.40	4.01 ± 0.41	6.96 ± 0.19
Paradiso	4.24 ± 0.08	30.68 ± 3.66	7.21 ± 0.53	4.25 ± 0.40	6.76 ± 0.15

reading of these texts and skills of the reader, especially short-term memory, modelled with three equivalent buffers, as I will show in Section 7.

Now, if the statistics of the three cantiche are different, what is the probability that a randomly selected canto from Divina Commedia could be confused/attributed to a cantica other than the one to which it belongs? In other words, what is the probability that a canto, belonging to *Inferno*, could be mathematically confused with a canto belonging to *Paradiso*, etc.? This analysis is interesting because it reveals how similar texts are to each other when considering one linguistic parameter at a time. In Section 6, I will present a geometric tool that serves the same purpose, but based on all linguistic parameters. The next section discusses this topic.

4 Probability density function of deep–language variables and likeness index

The probability density function (PDF) of a deep-language variable x can be modelled with a log-normal density function with three parameters [11, 17]:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma(x - x_o)} \exp\left\{-\frac{1}{2}\left[\frac{\log(x - x_o) - \mu}{\sigma}\right]^2\right\}, \quad x \geq x_o \tag{7}$$

In Eq.(7) the threshold is set at $x_o = 1$ for all deep-language variables. The mean value μ and the standard deviation σ (natural logs) are given by the linear mean value $m = \langle x \rangle$ and linear standard deviation s according to the following transformations [11, 17]:

$$\sigma^2 = \log\left[\left(\frac{s}{m-1}\right)^2 + 1\right] \tag{8}$$

$$\mu = \log(m-1) - \frac{\sigma^2}{2} \tag{9}$$

The linear mean and standard deviation are those reported in Table 4. Figures 1, 2, 3, 4 and 5 show the resulting PDF for each variable.

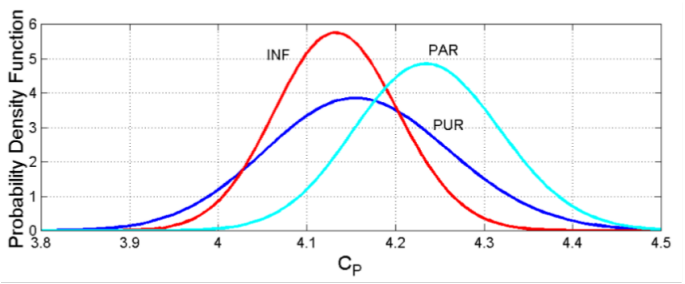

Figure 1. Probability density function, $f(x) = C_P$ for *Inferno* (red line), *Purgatorio* (blue line) and *Paradiso* (cyan line).

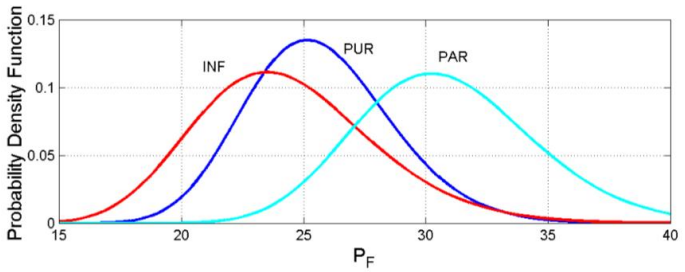

Figure 2. Probability density function, $f(x) = P_F$ for *Inferno* (red line), *Purgatorio* (blue line) and *Paradiso* (cyan line).

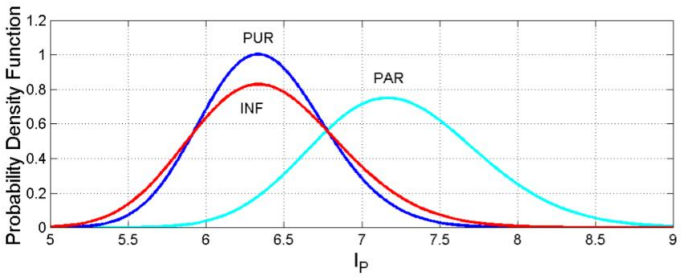

Figure 3. Probability density function, $f(x) = I_P$ for *Inferno* (red line), *Purgatorio* (blue line) and *Paradiso* (cyan line).

These PDFs are useful to assess whether a text can be confused with another text, for that particular deep-language variable. This assessment is possible by calculating the likeness index I_L given by [14]:

$$I_L = \left[\int_{x_{\min}}^{\infty} g_j(x)\,dx + \int_{-\infty}^{x_{\min}} g_k(x)\,dx\right] \tag{10}$$

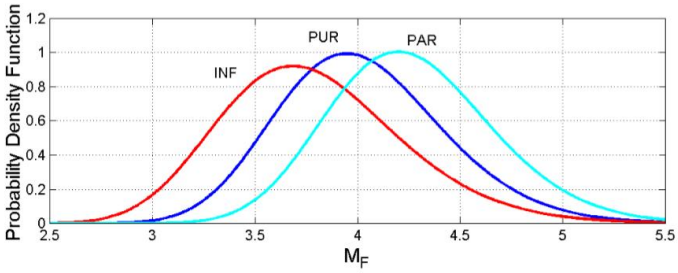

Figure 4. Probability density function, $f(x) = M_F$ for *Inferno* (red line), *Purgatorio* (blue line) and *Paradiso* (cyan line).

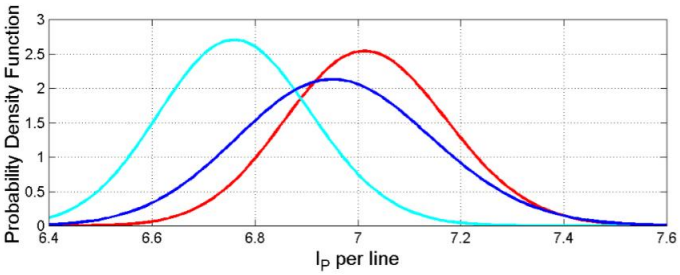

Figure 5. Probability density function, $f(x) = I_{P,l}$ for *Inferno* (red line), *Purgatorio* (blue line) and *Paradiso* (cyan line).

In Equation (10) $g_j(x)$ and $g_k(x)$ are the (log-normal) PDFs of text j and text k, for the same variable, like those shown in Figures 1, 2, 3, 4 and 5. The decision threshold $x_{\min}$ is given by the intersection between $g_j(x)$ and $g_k(x)$, a choice that minimizes the probability of error, as discussed in Reference [14]. The integral limits in Equation (10) assume $\mu_j < \mu_k$, therefore, $x_{\min} > \mu_j$.

The likeness index is bound in the range $0 \leq I_L \leq 1$. If $I_L = 0$, then there is no overlap between the two densities. Their mean values are centered at $-\infty$ and at $+\infty$, respectively, or the two densities have collapsed to Dirac delta functions. If $I_L = 1$, then the two densities are identical, i.e. text j and text k coincide. Tables 5, 6, 7 and 8 report I_L.

Table 5. Likeness index I_L for C_P in the three cantiche.

Cantica	*Inferno*	*Purgatorio*	*Paradiso*
Inferno	1	0.83	0.50
Purgatorio	0.83	1	0.66
Paradiso	0.50	0.66	1

From the PDFs shown in Figures 1, 2, 3, 4 and 5 and Tables 5, 6, 7 and 8, it is evident that, regardless of the deep–language variable, as we pass from *Inferno* to *Purgatorio* and to *Paradiso*, the likeness index decreases. In conclusion, this analysis, more detailed compared to the one that considers only mean values, confirms that *Inferno* and *Purgatorio* are mathematically more similar to each other than to *Paradiso*.

Table 6. Likeness index I_L for P_F in the three *cantiche*.

Cantica	*Inferno*	*Purgatorio*	*Paradiso*
Inferno	1	0.81	0.36
Purgatorio	0.81	1	0.43
Paradiso	0.36	0.43	1

Table 7. Likeness index I_L for I_P and $I_{P,\ell}$ (in parentheses) in the three *cantiche*.

Cantica	*Inferno*	*Purgatorio*	*Paradiso*
Inferno	1	0.95 (0.85)	0.41 (0.56)
Purgatorio	0.95 (0.85)	1	0.36 (0.57)
Paradiso	0.41 (0.56)	0.36 (0.57)	1

In the next section, I consider the difficulty that an educated reader may find in reading the cantiche. This difficulty can be relatively measured by a universal readability index.

5 Universal Readability Index

The universal readability index G_U measures the relative reading difficulty of texts in any alphabetical language [22]. Of course, it is understood that readers are enough educated to understand the texts. The index is given by:

$$G_U = 89 - 10 \times kC_P + \frac{300}{P_F} - 6 \times (I_P - 6) \qquad (11)$$

$$k = \frac{\langle C_{P,ITA} \rangle}{\langle C_{P,Lan} \rangle} \qquad (12)$$

In Eq.(12), the mean value $\langle C_{P,ITA} \rangle$ refers to the Italian language, $\langle C_{P,Lan} \rangle$ refers to the language of the text. By using Equations (11) and (12), the mean value $\langle kC_P \rangle$ is forced to be equal to that found in Italian. The rationale for this choice, discussed in detail in Reference [22], is that C_P is a parameter typical of a language which, if not scaled, would bias G_U without really quantifying the reading difficulty of readers, who in their own language are used 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 or longer than Italian. In this paper, since the texts are in Italian, $k = 1$.

Table 8. Likeness index I_L for M_F in the three cantiche.

Cantica	*Inferno*	*Purgatorio*	*Paradiso*
Inferno	1	0.76	0.55
Purgatorio	0.76	1	0.76
Paradiso	0.55	0.76	1

G_U is "decoded" as a readability index by considering the number of schooling years attended in Italy according to the chart shown in Figure 6. Now, since what matters is a relative comparison between the readability of the texts, the chart shown in Figure 6, although refers to the Italian school system, can nevertheless be used – I believe – as a reference in any language.

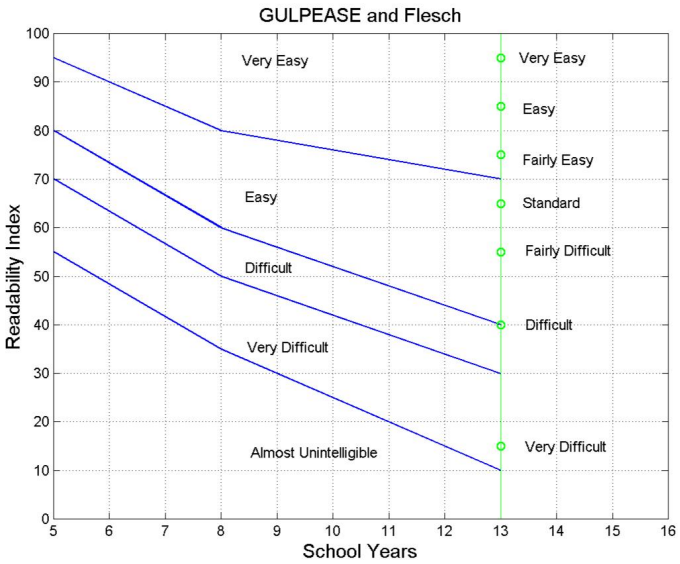

Figure 6. Chart for estimating the number of school years, for a given G_U, according to text reading difficulty as a function of the number of school years attended in Italy [22]. The continuous lines divide the quadrant in areas of similar performance of texts. In Italy, primary (elementary) school lasts 5 years, junior high school lasts 3 years, high school lasts 5 years. Pupils attend school until the age of 19. For comparison, the green vertical axis on the right refers to Flesch Reading Ease index [23, 24].

Eq.(11) and Figure 6 indicate that the larger G_U, the more readable the text is, for the same number of school years. The continuous lines divide the quadrant in areas of similar reading difficulty, such as "almost unintelligible", "very difficult", etc. For example, the area labelled "easy" indicates all combinations of values of G_U and school years required to declare a text "easy" to read for these readers. In all cases, the chart shows that the readability index that a reader can tolerate decreases as the number of school years increases.

Figures 7, 8 and 9 show the scatterplots G_U versus C_P, G_U versus I_P, and G_U versus P_F. Figure 7 shows that C_P plays a minor role in determining G_U [11], because the two variables are practically uncorrelated (correlation coefficients are reported in Table 9). More interestingly, Figures 8 and 9 show a clear inverse proportionality of G_U with I_P and P_F, and good correlation (Table 9). The distinction between *Inferno* and *Purgatorio* on the one hand and *Paradiso* on the other is always clear in any case.

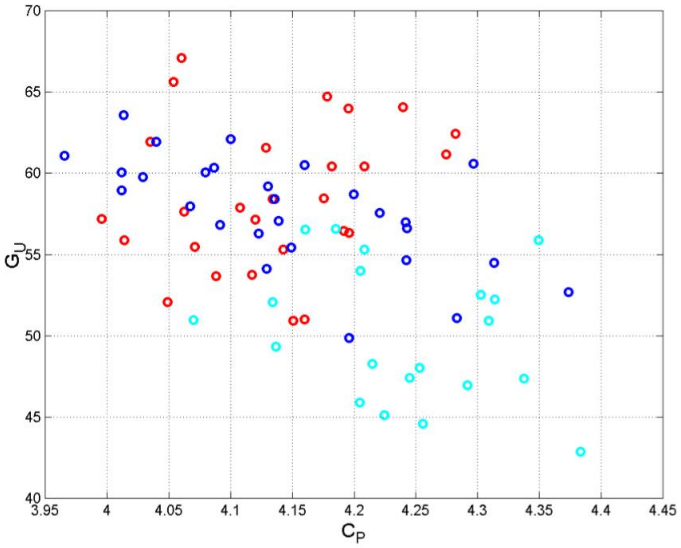

Figure 7. Scatterplot G_U versus C_P. *Inferno*, red circles; *Purgatorio*, blue circles; *Paradiso*, cyan circles.

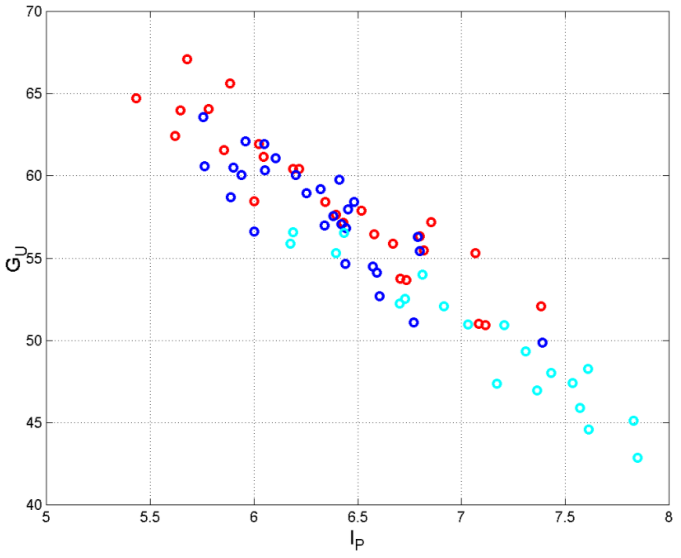

Figure 8. Scatterplot G_U versus I_P. *Inferno*, red circles; *Purgatorio*, blue circles; *Paradiso*, cyan circles.

Table 10 reports mean value and standard deviation of G_U for the three cantiche and the estimated number

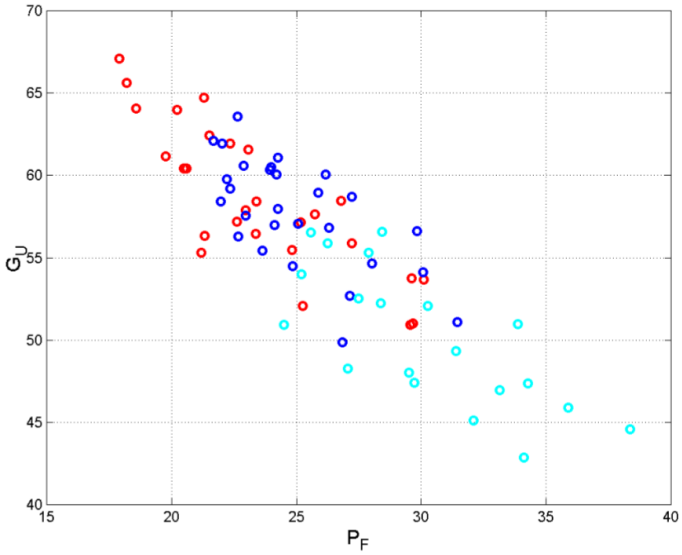

Figure 9. Scatterplot G_U versus P_F. *Inferno*, red circles; *Purgatorio*, blue circles; *Paradiso*, cyan circles.

Table 9. Correlation coefficients in scatterplots shown in Figures 7, 8 and 9, between the indicated variables.

Cantica	G_U versus C_P	G_U versus I_P	G_U versus P_F
Inferno	+0.210	-0.929	-0.830
Purgatorio	-0.633	-0.847	-0.663
Paradiso	-0.276	-0.957	-0.740

of years to make the cantiche "difficult" or " easy" to read.

For this variable, *Inferno* and *Purgatorio* show the same G_U and both cantiche require approximately the same schooling. *Paradiso* is decidedly less readable, and, for the same label –"difficult" or "easy"– requires more schooling.

In the Italian High Schools, the reading of specific canti of the Divina Commedia in three successive years is mandatory, starting from *Inferno*. Fortunately, this sequence agrees with the increase of schooling indicated in Table 10, since *Paradiso* is read in the last year of High School.

In the following, the analysis with PDFs of G_U confirms that *Inferno* and *Purgatorio* are mathematically more similar to each other than to *Paradiso*.

Table 10. Mean value and standard deviation of the universal readability index G_U in the three cantiche and the minimum number of school years (according to Figure 6) to define the text "difficult" or "easy" to read.

Cantica	G_U	**School Years**	
		Difficulty	**Easy**
Inferno	58.11 ± 4.17	6.5	8.5
Purgatorio	57.13 ± 3.78	6.7	8.7
Paradiso	49.26 ± 3.95	8.5	10.5

For a wide range, the PDF of $x = G_U$ can be modelled with a Gaussian density function because the probability of $G_U < 0$ (negative values are not decoded according to Figure 6) is negligible, because $\langle G_U \rangle >> 0$. The model is the following:

$$f(x = G_U) = \frac{1}{\sqrt{2\pi}s} \exp\left\{-\frac{1}{2}\left[\frac{x-m}{s}\right]^2\right\} \qquad (13)$$

The mean value m and the standard deviation s are given in Table 10. Figure 10 shows the PDFs of the three cantiche and Table 11 reports the likeness index. It is confirmed that *Inferno* and *Purgatorio* are mathematically more similar to each other than to *Paradiso*. The probability of confusing *Inferno* or *Purgatorio* with *Paradiso* is very small.

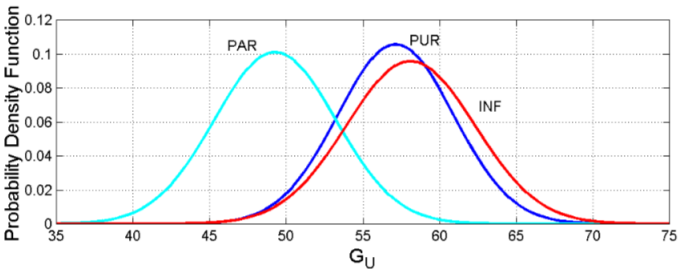

Figure 10. Probability density function, $f(x) = G_U$ for *Inferno* (red line), *Purgatorio* (blue line) and *Paradiso* (cyan line).

Table 11. Likeness index I_L for G_U in the three cantiche.

Cantica	***Inferno***	***Purgatorio***	***Paradiso***
Inferno	1	0.90	0.28
Purgatorio	0.90	1	0.31
Paradiso	0.28	0.31	1

Texts can also be represented geometrically [11], a visualization that immediately shows their mathematical similarity or "closeness." This analysis is more compact than the analysis involving single linguistic variables because it considers them jointly, as illustrated in the next section.

6 Geometric representation of alphabetic texts

To place the cantiche studied here within the framework of Italian literature, I have considered some other masterworks of the Italian literature, namely:

Fermo e Lucia and *I promessi sposi* (*The Betrothed*) written by Alessandro Manzoni (XIX century); *Decamerone* written by Giovanni Boccaccio (published 1353); *Canzoniere* (1336–1374) and *Trionfi* (1351) written by Francesco Petrarca. The first three texts are novels, the last two are poems. Even more interestingly, Trionfi and Canzoniere were written only a few years after the Divina Commedia and, furthermore, Trionfi was written in Dante's tercets, so its mathematical structure should be very similar to that of the Divina Commedia and should, therefore, allow for a meaningful comparison. Table 12 lists the average values of the linguistic variables of these texts.

Table 12. Mean value of the deep-language variables in the indicated literary texts.

Text	$\langle C_P \rangle$	$\langle P_F \rangle$	$\langle I_P \rangle$	$\langle M_F \rangle$
I promessi sposi	4.60	24.83	5.30	4.63
Fermo e Lucia	4.75	30.98	7.17	4.30
Decamerone	4.48	44.27	7.79	5.69
Canzoniere	4.41	31.28	6.48	4.82
Trionfi	4.39	30.37	6.62	4.59

The mean values of Tables 4 and 12 can be used to model the texts as vectors in the first Cartesian quadrant, a modeling discussed in detail in Ref. [11] and briefly recalled here. This geometric representation uses the four linguistic variables simultaneously to synthesize the mathematical characteristics of a text and compare it with other texts.

Let us consider the following six vectors of the indicated components of deep-language variables:

$$\begin{aligned} \overrightarrow{R_1} &= (\langle C_P \rangle, \langle P_F \rangle), \quad \overrightarrow{R_2} = (\langle M_F \rangle, \langle P_F \rangle), \\ \overrightarrow{R_3} &= (\langle I_P \rangle, \langle P_F \rangle), \quad \overrightarrow{R_4} = (\langle C_P \rangle, \langle M_F \rangle), \\ \overrightarrow{R_5} &= (\langle I_P \rangle, \langle M_F \rangle), \quad \overrightarrow{R_6} = (\langle I_P \rangle, \langle C_P \rangle) \end{aligned}$$

and their resulting vector sum:

$$\overrightarrow{R} = \sum_{k=1}^{6} \overrightarrow{R_k} = \overrightarrow{x_i} + \overrightarrow{y_j} \tag{14}$$

Now, two texts are likely mathematically connected – they show close ending points of vector (14) – if their relative Pythagorean distance is small.

By considering the vector components x and y of Eq. (14), Figure 11 shows the scatterplot where X and Y are normalized coordinates calculated by setting *Inferno* (INF) at the origin ($X = 0, Y = 0$) of the Cartesian quadrant, and *Paradiso* (PAR) at ($X = 1, Y = 1$), according to the linear transformation:

$$X = \frac{x - x_{INF}}{x_{PAR} - x_{INF}} \tag{15}$$

$$Y = \frac{y - y_{INF}}{y_{PAR} - y_{INF}} \tag{16}$$

These linear transformations are very useful because they allow to zoom in on a specific geometric area to observe and study the relationship between texts.

From Figure 11, we can observe the following interesting features:

- *Inferno* and *Purgatorio* are very close to each other.
- *Paradiso* is clearly displaced from *Inferno* and *Purgatorio*.
- *Trionfi* and *Canzoniere* overlap almost completely and are very close to *Paradiso*.
- Mathematically, Petrarch writes as Dante does in *Paradiso*, not as Dante does in *Inferno* and *Purgatorio*.
- The readability index of the texts is consistent with their geometric position: adjacent texts have similar readability indices. Moreover, the closer the texts are to the origin $(0, 0)$, the greater their readability index.

It is very interesting to notice that although *Trionfi* was written in Dante's tercets and *Canzoniere* was not, the two texts mathematically coincide. In other words, they can be considered as a mathematical "imprint" of Petrarca's writings, regardless of the rhyme used.

Indeed, although *Trionfi* was written in Dante's tercets, its mathematical structure is practically identical to that of *Canzoniere*, which is not written in tercets. Therefore, the particular rhyme and hendecasyllabic verses cannot be used to distinguish the two poetic texts. In other words, the structure of chained hendecasyllabic verses does not affect on the deep-language variables, but only on the word interval per line, $I_{P,\ell}$. More on $I_{P,\ell}$ in the next section.

Finally, it can be noted that the vector representation of texts can highlight, as already observed in [11], the improvement that an author can make to his own text, after a long revision. This is the case of the two novels written by Alessandro Manzoni, in which *The Betrothed* is actually the "revised" version of *Fermo and Lucia*, written after about twenty years. In Figure 11 they

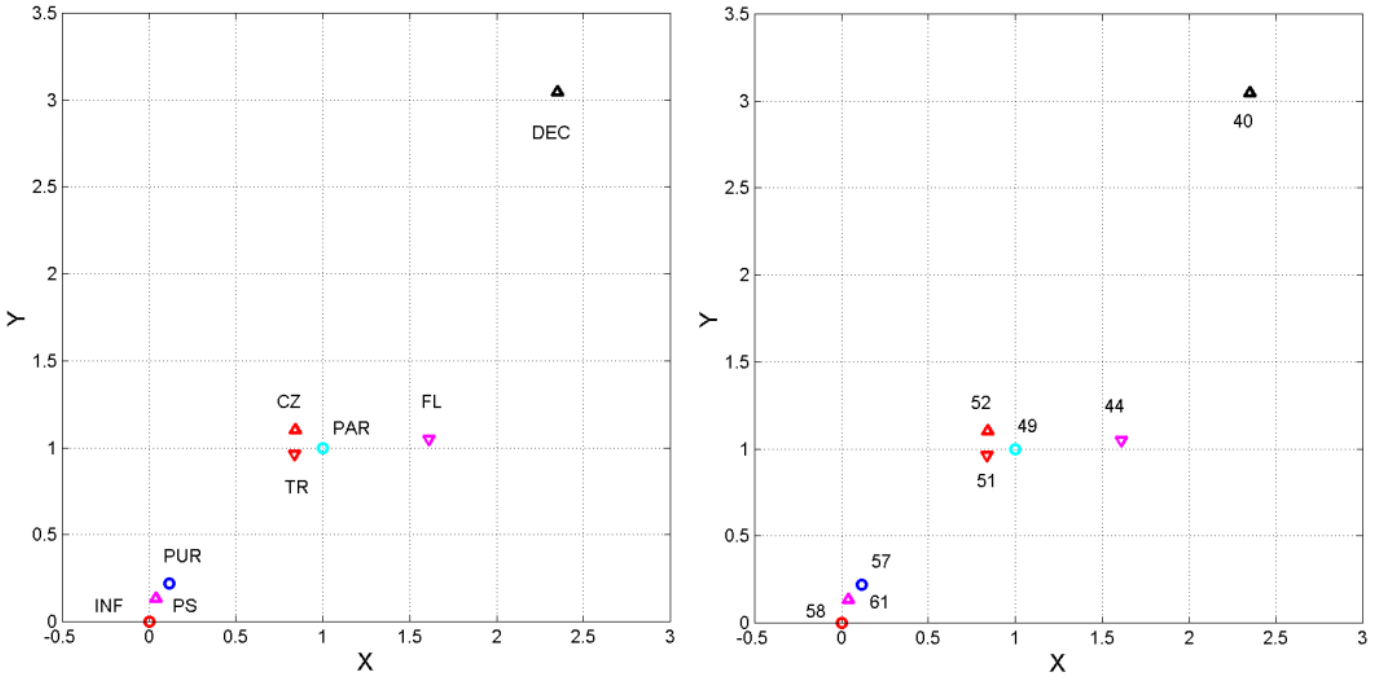

Figure 11. Left panel: Scatterplot of normalized coordinates X and Y of the ending point of vector (14) such that *Inferno*, INF, red circle is at (0,0) and *Paradise*, PAR, cyan circle, is (1,1). *Purgatorio*, PUR, blue circle; *Canzoniere*, CZ, upward red *triangle*; Trionfi, TR, downward red triangle; *Decamerone*, upward black triangle; I promessi sposi, PS, upward magenta triangle; Fermo e Lucia, downward magenta triangle. Right: Universal readability index of the texts. Note that the closer the text is to the origin of the Cartesian axes, the greater its readability index.

are clearly each other displaced and have different readability indices.

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 4, do introduce data scattering, therefore now I study and discuss this issue by calculating the probability (called again "error" probability) that a text may be mathematically confused with another one, as done in Ref. [25].

Besides the vector $\vec{R}$ of Eq. (14) – 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 $\vec{R}$. In this case, the final random vector describing a text is given by:

$$\vec{T} = \vec{R} + \vec{\rho} \tag{17}$$

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-language variables that determine the components x and y of $\vec{R}$ – let their sum 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:

$$\rho = \sigma_\rho \tag{18}$$

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 [25].

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. As discussed in [25], the hypothesis of substantial uniformity around $\vec{R}$ should be justified within 1-sigma bounds given by Eq.(18).

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. Since probabilities are proportional to areas, we get the following relationships:

$$\frac{A}{A_1} = \frac{P(A_1, A_2)}{P(A_1)} = \frac{P(A_2 \mid A_1)P(A_1)}{P(A_1)} = P(A_2 \mid A_1) \tag{19}$$

$$\frac{A}{A_2} = \frac{P(A_1, A_2)}{P(A_2)} = \frac{P(A_1 \mid A_2)P(A_2)}{P(A_2)} = P(A_1 \mid A_2) \tag{20}$$

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

A synthetic parameter which highlights how much two texts can be erroneously confused with each other is the average conditional probability:

$$p_e = P(A_2 \mid A_1)P(A_1) + P(A_1 \mid A_2)P(A_2) \tag{21}$$

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

$$p_e = 0.5 \times \left[P(A_2 \mid A_1) + P(A_1 \mid A_2)\right] \tag{22}$$

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, therefore there is a deterministic mathematical connection involving the deep-language parameters.

Figure 12 shows the result of this investigation on the three *cantiche* in the normalized variable plane of Figure 11.

From Figure 12 it is evident that $p_e = 0$ between *Inferno*, or *Purgatorio*, and *Paradiso* (no overlapping). On the contrary $p_e = 0.70$ between *Inferno* and *Purgatorio*. This result confirms the analysis of the individual parameters in Section 5 and reiterates that *Inferno* and *Purgatorio* are mathematically very similar and substantially different from *Paradiso*.

The difficulty of reading a text, measured by G_U, is related to the reader's and writer's short-term memory. The next section addresses this important topic.

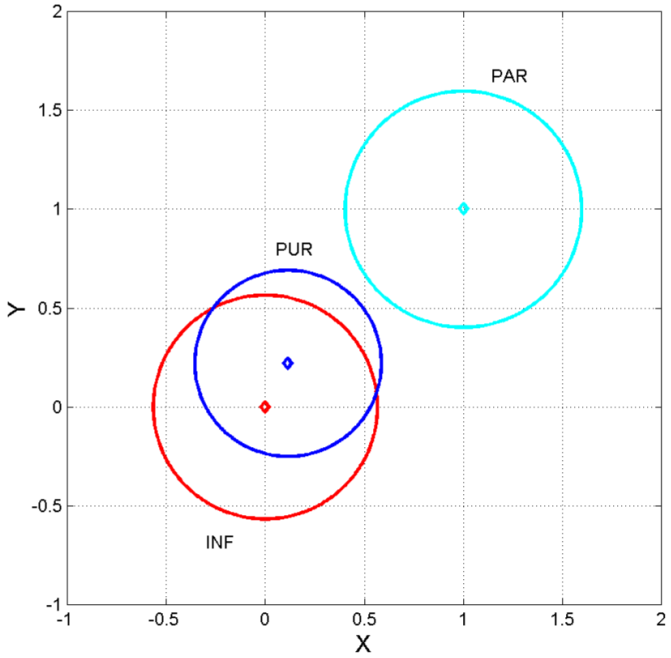

Figure 12. Scatterplot of normalized coordinates X and Y of the ending point of vector (17) such that *Inferno*, INF, red circle is at (0,0) and *Paradise*, PAR, cyan circle, is (1,1). *Purgatorio*, PUR, blue circle. The red, blue and cyan circles refer to 1-sigma contour lines.

7 Equivalent buffers of the extended short-term memory

To memorize a sentence, the human short-term memory can be modelled by three equivalent buffers in series, Figure 13 [26]. The first buffer is linked to the number of syllables/characters per word. The second buffer is linked to the number of words between two consecutive interpunctions - i.e., the word interval IP, which follows Miller's 7±2 Law [27]. As for the many studies on the short-term memory and its several modelling, the reader is referred, among others, to classic theoretical accounts of the construct and their later reassessments [35–37]; to the empirical debate on the precise size of the capacity limit that followed Miller's original proposal [30–32]; to chunking- and compression-based explanations of memory span [28, 29, 34, 45]; to developmental and individual-differences perspectives, including their specific relevance to reading [38–41]; to neuroscientific and conceptual reviews of the construct [42, 43, 46]; and to critical reassessments and studies specifically concerned with working memory during text comprehension [33, 44]. The third buffer is linked to the number of word intervals contained in a sentence, described by the variable M_F. I refer to the overall model as the extended short-term memory (E-STM), not to be confused with the so-called working

memory of cognitive scholars.

Indeed, note that this diagram shows the capacity of the buffers and how they are filled, not the temporal sequence in which the mind actually processes textual information, a process largely unknown even to cognitive scientists. In other words, Figure 13 shows an "equivalent" flow of information across the three buffers, not a precise temporal correspondence which is largely unknown.

The features of the three buffers of Figure 13 applicable to the Divina Commedia are discussed in the following subsections.

7.1 First E-STM buffer: syllables and characters

As for the first buffer, Figures 14, 15 and 16 show the scatterplots between characters and syllables, characters and words and syllables and words. Table 13 reports slope and correlation coefficient of the regression lines.

From Figures 14, 15 and 16 and Table 13, we can notice that for these relationships there are no striking differences between the three *cantiche*. In other words, the first buffer stores syllables and characters the same way. At this level, higher processing concerning the meaning has not fully yet taken place. See also the poor correlation of C_P with G_U in Figure 7. The higher cognitive processing involving meaning seem to occur in the equivalent second buffer – which processes the words contained in a word interval – and in the equivalent third buffer, which processes the word intervals contained in a sentence.

7.2 Second E–STM buffer: I_P and $I_{P,l}$

As for the second buffer, Figure 17 shows I_P versus P_F. The trend is typical of alphabetical languages [12, 47] and confirms two fundamental facts:

- I_P ranges in Miller's bounds, therefore this variable describes Miller's law in alphabetic texts.
- If P_F increases, I_P does also increase but not linearly, because the second buffer cannot hold, approximately, a number of words greater than Miller's Law upper bound, therefore saturation must occur.

Notice that *Inferno* and *Purgatorio* are very similar and different from *Paradiso*.

The non-linear best fit drawn in Figure 17 can be modelled by the relationship [12]:

$$I_P = (I_{P\infty} - 1) \times \left[1 - e^{-\frac{(P_F - 1)}{(P_{F0} - 1)}}\right] + 1 \qquad (23)$$

Table 14 reports the constants present in Eq. (23).

From Table 14 and Figure 17, we can notice that the *asymptote* value $(I_{P\infty} - 1)$ is 6.50 words for *Inferno*, 6.20 words for *Purgatorio* and 7.96 words for *Paradiso*. Notice that the *asymptotes* of *Inferno* and *Purgatorio* are almost at the center of Miller's range while that of *Paradiso* is near the upper bound. The constant $(P_{F\infty} - 1)$ at which the exponential becomes e^{-1} is 11.84 words in *Inferno*, 11.05 in *Purgatorio* and 18.35 in *Paradiso*. These constants mark the difference between *Paradiso* and the other two *cantiche*. In *Paradiso saturation* occurs for larger values of P_F, therefore for longer sentences.

It is very interesting to note that the word interval per line $I_{P,\ell}$, is about 7 in the three *cantiche* (Table 4), with very small standard deviation, and that this value is precisely the central value of Miller's law.

The fact that one line is clearly distinct from the next seems to engage the E–STM as does the word interval – words between two successive *interpunctions* – but with a fixed number of words. In other terms, *hendecasyllabic* verses must have arisen spontaneously in the poets' minds because of Miller's law. Although *hendecasyllables* are theoretically based on syllable counting – as any scholar of literature would assume – they were likely also created unconsciously because they match the central value of Miller's 7 ± 2 Law. Therefore, I hypothesize that *hendecasyllabic* verses are easy to memorize, not only because of the distinctive structure mentioned in Section 2, but also because they fit Miller's Law.

7.3 Third E–STM buffer: sentences

As for the third buffer, Figure 18 shows M_F versus P_F in the *Divina Commedia* and several novels of the Italian and English *literatures* (for these data see Reference [48]). Table 15 reports the slope and correlation coefficient of the regression lines of the data referring to the Italian and English *literatures*.

The usual identity between *Inferno* and *Purgatorio* is observed, as also the difference with *Paradiso*, which requires larger buffers. In reading *Paradiso*, a more powerful third buffer is required than in reading *Inferno* or *Purgatorio*.

From Figure 18 and Table 15 we can notice that Italian and English literatures show, practically, the same regression line. The identical relationship between

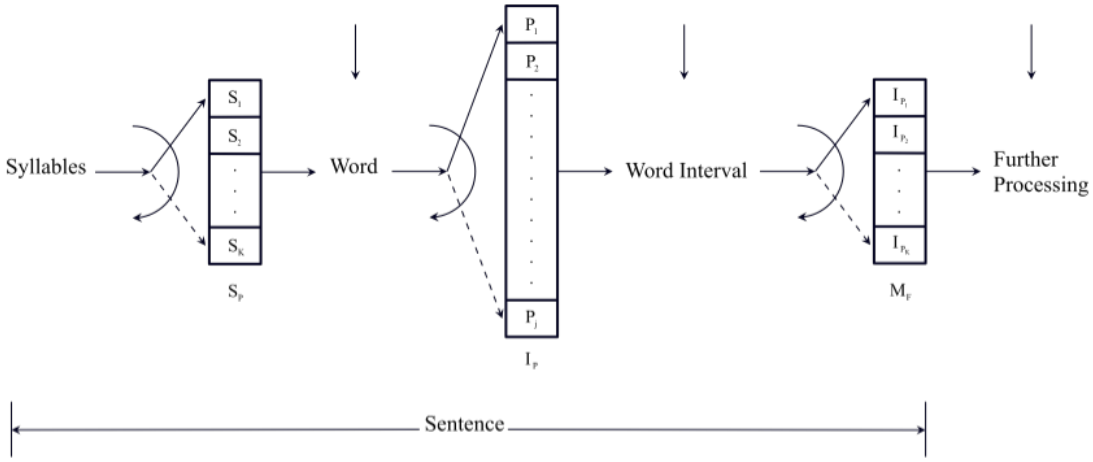

Figure 13. Modelling of the Extended Short-term Memory (E-STM) with three equivalent buffers in series [26]. Syllables $S_1, S_2, \ldots, S_k$ are stored in the first processor, from 1 to about 3–4 items, until a space or an interruption is introduced (vertical arrow) to fix the length of the word. Words $P_1, P_2, \ldots, P_j$ are stored in the second processor – approximately in Miller's range – until an interruption (vertical arrow) is introduced to fix the length of I_P. The word interval I_P is then stored in the third processor, from about 1 to 6 items, until the sentence ends with a full stop, a question mark or an exclamation mark (vertical arrow). The process is then repeated for the next sentence. The first buffer is linked to the number of syllables/characters per word. The second buffer is linked to the number of words between two consecutive interruptions, I_P. The third buffer is linked to the number of word intervals contained in a sentence, described by the variable M_F.

Table 13. Slope and correlation coefficients in scatterplots shown in Figures 14, 15 and 16, between the indicated variables.

Cantica	Characters vs Syllables		Characters vs Words		Syllables vs Words	
	Slope	Corr. Coeff.	Slope	Corr. Coeff.	Slope	Corr. Coeff.
Inferno	2.45	0.9866	4.13	0.9625	1.69	0.9326
Purgatorio	2.44	0.9901	4.14	0.8814	1.70	0.8802
Paradiso	2.43	0.9878	4.24	0.9017	1.75	0.8983

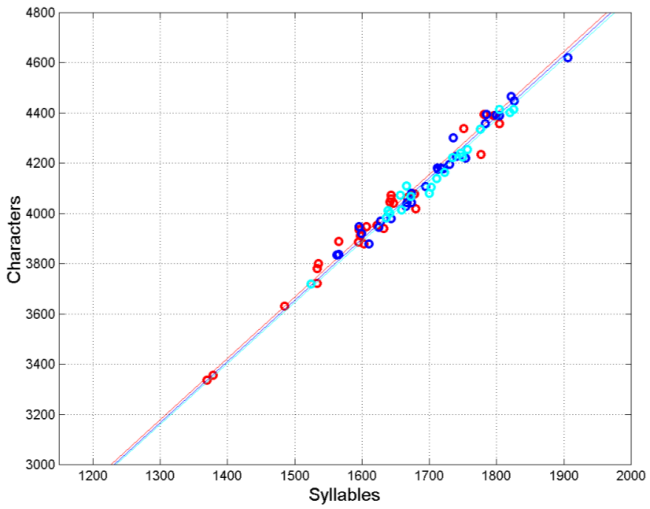

Figure 14. Scatterplot of characters versus syllables. Inferno, red circles; Purgatorio, blue circles; Paradiso, cyan circles.

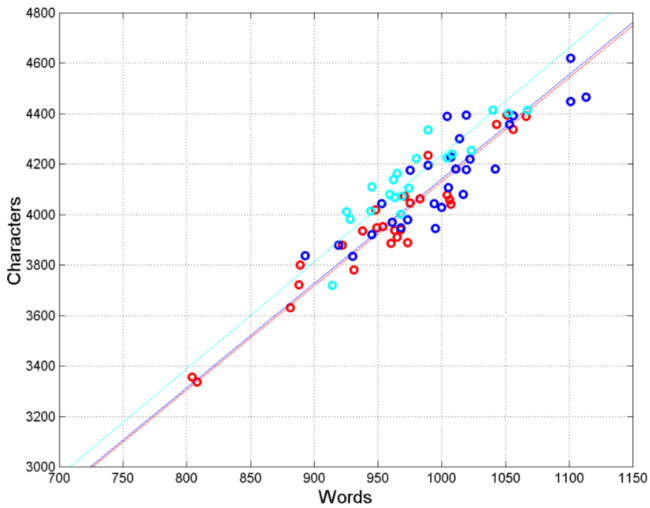

Figure 15. Scatterplot of characters versus words. Inferno, red circles; Purgatorio, blue circles; Paradiso, cyan circles.

M_F and P_F indicates that these variables describe an equivalent third buffer that, for a given sentence length, has the same capacity, expressed by M_F, regardless of language. In other words, the linear relationship given in Table 15 is typical of alphabetical languages and the human mind.

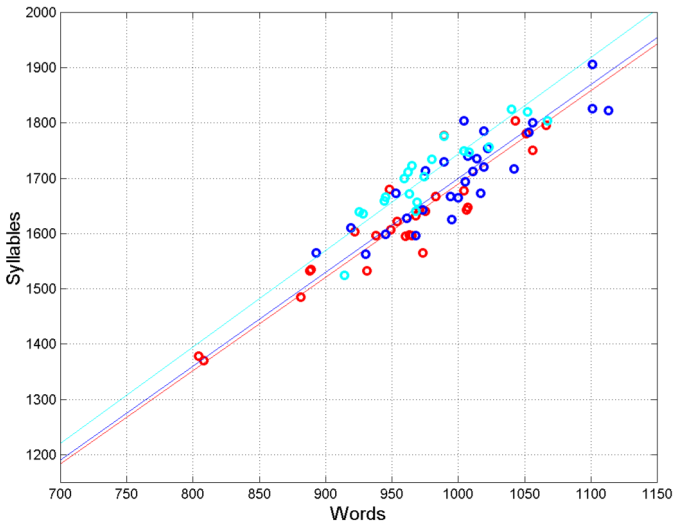

Figure 16. Scatterplot of syllables versus words. Inferno, red circles; Purgatorio, blue circles; Paradiso, cyan circles.

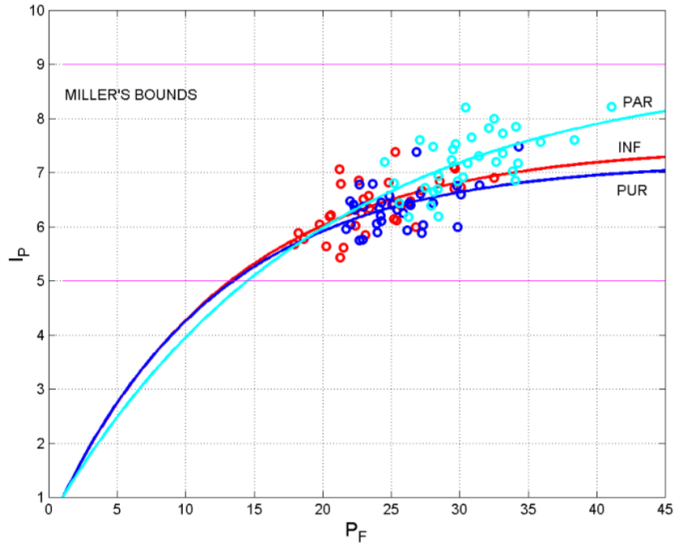

Figure 17. I_P versus P_F. *Inferno* (INF), red circles and red line; *Purgatorio* (PUR), blue circles and blue line; *Paradiso* (PAR), cyan circles and cyan line. Miller's range is between $I_P = 5$ and $I_P = 9$.

8 Summary, future work and conclusion

Dante Alighieri's *Divina Commedia* is the preeminent work of Italian literature and one of the greatest works of world literature. At each stage of the journey through *Inferno*, *Purgatorio* and finally *Paradiso*, Dante adapts his language to the different contexts he describes. Depending on the characters he converses with, his vocabulary ranges from everyday language to scientific, from the popular to the refined, used by the most cultured people. Dante's language and narrative register evolve as his journey towards *Paradiso* progresses.

Indeed, Dante achieved a literary revolution, bringing

Table 14. Values of the parameters of Eq. (23) for the indicated *cantica*.

Cantica	$I_{P\infty}$	P_{F0}
Inferno	7.50	12.84
Purgatorio	7.20	12.05
Paradiso	8.96	19.35

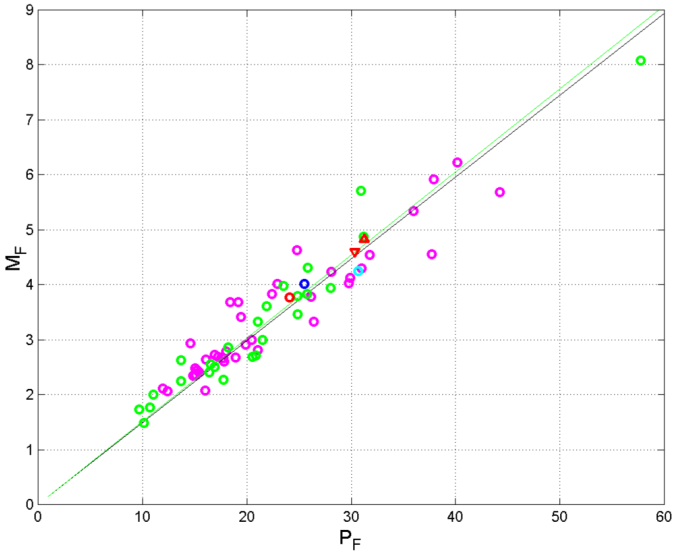

Figure 18. Scatterplot of M_F versus P_F. Italian literature: magenta circles and black line; English literature: green circles and green line. *Inferno*, red circle; *Purgatorio*, blue circle; *Paradiso*, cyan circle. *Canzoniere*, upward red triangle; *Trionfi*, downward red triangle.

together in a single great work the various types of language that had previously been specific to comedy or tragedy. This is why scholars, referring to the *Divina Commedia*, speak of *pluristylism*. Compared to *Inferno* and *Purgatorio*, the language of *Paradiso* clearly increases in complexity, with a quest for absolute linguistic perfection to express profoundly significant theological concepts. The reading thus becomes more demanding and difficult to interpret.

Since unanimously scholars do mention large differences between *Paradiso* and the other two *cantiche* in many aspects, in this article I have examined whether the deep-language mathematical structure, of which Dante is unaware, is different in *Inferno*, *Purgatorio*, and *Paradiso*. The answer is affirmative, as I have shown in detail in the preceding sections.

In conclusion, the multidimensional analysis of the *Divina Commedia* has highlighted that:

- *Inferno* and *Purgatorio* are very similar and markedly different from *Paradiso*, in agreement with scholars of Italian Literature.

Table 15. Slope and correlation coefficient of the regression line M_F versus P_F of Italian and English literatures shown in Figure 18.

Literature	Slope	Correlation coefficient
Italian	0.149	0.937
English	0.151	0.969

- *Inferno* and *Purgatorio* are more readable than *Paradiso*, according to a universal readability formula.
- *Inferno* and *Purgatorio* require less powerful short-term memory equivalent buffers than *Paradiso*.
- The approach of scholars using the traditional tools and the approach of scientists using mathematical tools and a multidimensional analysis can reinforce each other in making comparisons between texts written by different authors or by the same author.
- The origin of the popular *hendecasyllabic* verse, used by Dante and other authors, most likely was not devised only by counting syllables, but also, unconsciously, by counting words, since any verse contains 7 words, the central value of Miller's 7 ± 2 Law.

My conjecture that *hendecasyllables* are due more to short-term memory than to the particular structure of the verse is supported by a study I am currently conducting: specifically, hexameters from Greek (e.g., *Iliad*, *Odyssey*) and Latin literature (e.g., *Aeneid*, *De Rerum Natura*) consistently indicate that each hexameter contains about seven words, with a negligible standard deviation. In other words, although hexameters and *hendecasyllables* were structured according to different parameters and historical developments, each always contains approximately seven words. This result, in my opinion, confirms that the basic structure of these verses is due to the limitations of the short-term memory. Recall that in ancient times these verses were largely memorized. I will present these results in a future article.

In conclusion, further studies could be done on other poems and other structure of verse, written in any alphabetic language and in any period, to assess their underlying mathematical structure and determine whether it is similar to that found in the *Divina Commedia*.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

Emilio Matricciani served as an Editor-in-Chief of the *Journal of Mathematical Studies of Literature* at the time of manuscript submission. To ensure the integrity of the peer-review process, Emilio Matricciani was not involved in the editorial handling, peer review, or decision-making process for this manuscript, which was handled independently by another editor.

AI Use Statement

The author declares that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] De Sanctis, F. (2006). *Storia della letteratura italiana* (First published 1870). Rizzoli.

[2] Sapegno, N. (1968). *La Divina Commedia*. La Nuova Italia.

[3] Porena, M. (1969). *La Divina Commedia*. Zanichelli.

[4] Bosco, U., & Reggio, G. (2002). *La Divina Commedia*. Le Monnier.

[5] Cella, R. (2015). *Storia dell'italiano*. Società Editrice Il Mulino.

[6] Langella, G., Frare, P., Gresti, P., & Motta, U. (2019). *Amor mi mosse. Letteratura italiana. L'instaurazione del canone. I nuovi classici. Dalle origini all'età comunale*. Bruno Mondadori.

[7] Ferroni, G. (2021). Storia della letteratura italiana. *Dalle origini al Quattrocento*. Mondadori Università.

[8] Carrai, S., & Inglese, G. (2023). *La letteratura italiana del Medioevo*. Carocci.

[9] Frare, P., & Brenna, S. (2023). *Dalle origini a Leopardi*. La letteratura italiana e le sue grandi opere. Pearson.

[10] Martin, M. (1995). *A Linguistic History of Italian*. Routledge.

[11] 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*(3), 373–406. [CrossRef]

[12] Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint

on language. *Behavioral and brain sciences, 39*, e62. [CrossRef]

[13] Burrows, J. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and linguistic computing, 17*(3), 267-287. [CrossRef]

[14] Matricciani, E. (2022). Linguistic mathematical relationships saved or lost in translating texts: Extension of the statistical theory of translation and its application to the new testament. *Information, 13*(1), 20. [CrossRef]

[15] Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. Wiley, New York. http://andrei.gorea.free.fr/Teaching_fichiers/SDT%20and%20Psytchophysics.pdf

[16] Rybicki, J. (2012). The great mystery of the (almost) invisible translator. *Quantitative Methods in Corpus-Based Translation Studies: A practical guide to descriptive translation research, 231*, 231-248. https://www.torrossa.com/en/resources/an/5016442#page=242

[17] Crow, E. L., & Shimizu, K. (Eds.). (2018). *Lognormal distributions: Theory and applications*. Routledge.

[18] Robey, D. (1993). Scanning dante's the Divine Comedy. A Computer based Approach. *Literary and Linguistic Computing, 8*(2), 81-84. [CrossRef]

[19] Dakamsih, N. J. (2025). Threeness and the structure of Dante Alighieri's "Divina Commedia". *International Journal of Multidisciplinary Research and Growth Evaluation, 6*(1), 2105–2112. https://www.allmultidisciplinaryjournal.com/article/3629/threeness-and-the-structure-of-dante-alighieri-s-divine-comedy

[20] Perveen, T., Munawar, H., Zahra, M., & Lodhi, M. A. (2025). Corpus stylistic exploration of lexical patterns, semantic fields, and contextual meaning in Dante's Divina Commedia. *Contemporary Journal of Social Science Review, 3*(4), 114–131. [CrossRef]

[21] de Callataÿ, G., & de Callataÿ-van der Mersch, C. (2026). The Secret Diagrams of Dante's Divine Comedy. In *Lectori vago. Manuscripts, Libraries, and Classical Scholarship from the Middle Ages to the Early Modern Period* (pp. 95-113). Brepols Online. [CrossRef]

[22] Matricciani, E. (2023). Readability indices do not say it all on a text readability. *Analytics, 2*(2), 296-314. [CrossRef]

[23] Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology, 32(3), 221–233. [CrossRef]

[24] Flesch, R. (1974). The art of readable writing (Rev. and enlarged ed.). Harper & Row. https://archive.org/details/in.ernet.dli.2015.275839/page/n1/mode/2up

[25] Matricciani, E. (2024). Multi-dimensional data analysis of deep language in J.R.R. Tolkien and C.S. Lewis reveals tight mathematical connections. *AppliedMath, 4*(3), 927–949. [CrossRef]

[26] Matricciani, E. (2025). Equivalent processors modelling the short-term memory. *International Journal of Pure and Applied Mathematics Research, 5*(2), 1–28. [CrossRef]

[27] Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81–97. [CrossRef]

[28] Crowder, R. G. (1993). Short-term memory: Where do we stand?. *Memory & Cognition, 21*(2), 142-145. [CrossRef]

[29] Lisman, J. E., & Idiart, M. A. P. (1995). Storage of 7 ± 2 short-term memories in oscillatory subcycles. *Science, 267*(5203), 1512–1515. [CrossRef]

[30] Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences, 24*(1), 87-114. [CrossRef]

[31] Bachelder, B. L. (2001). The magical number 4= 7: Span theoryon capacity limitations. *Behavioral and Brain Sciences, 24*(1), 116-117. [CrossRef]

[32] Saaty, T. L., & Ozdemir, M. S. (2003). Why the magic number seven plus or minus two. *Mathematical and computer modelling, 38*(3-4), 233-244. [CrossRef]

[33] Richardson, J. T. (2007). Measures of short-term memory: a historical review. *Cortex, 43*(5), 635-650. [CrossRef]

[34] Mathy, F., & Feldman, J. (2012). What's magic about magic numbers? Chunking and data compression in short-term memory. *Cognition, 122*(3), 346-362. [CrossRef]

[35] Melton, A. W. (1963). Implications of short-term memory for a general theory of memory. *Journal of Verbal Learning and Verbal Behavior, 2*(1), 1–21. [CrossRef]

[36] Atkinson, R. C., & Shiffrin, R. M. (1971). The control of short-term memory. *Scientific american, 225*(2), 82-91. http://www.jstor.org/stable/24922803

[37] Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of verbal learning and verbal behavior, 14*(6), 575-589. [CrossRef]

[38] Case, R., Kurland, D. M., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of experimental child psychology, 33*(3), 386-404. [CrossRef]

[39] Grondin, S. (2001). A temporal account of the limitedprocessing capacity. *Behavioral and Brain Sciences, 24*(1), 122-123. [CrossRef]

[40] Pothos, E. M., & Juola, P. (2001). Linguistic structure and short term memory. *Behavioral and Brain Sciences, 24*(1), 138-139. [CrossRef]

[41] Conway, A. R., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence, 30*(2), 163-183. [CrossRef]

[42] Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., & Moore, K. S. (2008). The mind

and brain of short-term memory. *Annual Review of Psychology, 59*(1), 193–224. [CrossRef]

[43] Potter, M. C. (2012). Conceptual short-term memory in perception and thought. *Frontiers in Psychology, 3*, 113. [CrossRef]

[44] Jones, G., & Macken, B. (2015). Questioning short-term memory and its measurement: Why digit span measures long-term associative learning. *Cognition, 144*, 1-13. [CrossRef]

[45] Chekaf, M., Cowan, N., & Mathy, F. (2016). Chunk formation in immediate memory and how it relates to data compression. *Cognition, 155*, 96-107. [CrossRef]

[46] Norris, D. (2017). Short-term memory and long-term memory are still different. Psychological bulletin, 143(9), 992. [CrossRef]

[47] Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences, 108*(9), 3526-3529. [CrossRef]

[48] Matricciani, E. (2024). A Mathematical Structure Underlying Sentences and Its Connection with Short–Term Memory. *AppliedMath, 4*(1), 120-142. [CrossRef]

Appendix

A Raw textual data

The raw textual data I have used in this paper were extracted from the best available digital texts of the mentioned novels and poems, see https://liberliber.it/autori/autori-a/dante-alighieri/la-divina-commedia-edizione-petrocchi/ (last access April 7, 2026). The processing, although time consuming, consists of two successive steps:

1. Extraction of the raw data in each canto: number of syllables, characters, words, sentences, interpunctions; of course, the procedure was repeated 100 times for the 100 canti.

2. Further processing to calculate the deep–language parameters, their statistics and relationships.

For the first step, I have not used specialized software but tools that are available freely and known to any reader/writer, namely WinWord, with which I have counted, for each canto, the number of characters, words, sentences, and interpunctions. I have counted the number of syllables by using the software developed in https://www.separarensilabas.com/index\T1\textendashit.php (last access, April 7, 2026).

Before any processing, I have manually deleted titles, footnotes and other extraneous material present in the digital texts, leaving only the plain text.

For each text block (a canto in our case) WinWord directly provides the number of characters and words. For the other raw data, the count is very simple. The number of sentences is calculated in three steps: first, by replacing every full stop with a full stop; this action does not change the text but gives the number of substitutions, therefore the number of full stops. The same procedure is then repeated for question marks (second step) and exclamation marks (third steps). 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 number of sentences gives the number of interpunctions.

For the second step, I have developed software codes for each further processing, including scatterplots, regression lines, and so on.

B List of mathematical symbols and meaning

Symbol	Definition
C_P	Characters per word
G_U	Universal readability index
I_L	Likeness index
I_P	Word interval
$I_{P,\ell}$	Word interval per line
M_F	Word intervals per sentence
n_C	Number of characters
n_W	Number of words
n_S	Number of sentences
n_I	Number of interpunctions
P_F	Words per sentence

Emilio Matricciani, Professor of Telecommunications at Politecnico di Milano (retired 2022), Milan, Italy, taught Information Theory, Probability, Satellite and Terrestrial Communication Systems to graduate students. In addition to these institutional he still teaches Scientific Writing to PhD students Università La Sapienza, Rome, , and Università Federico II, Naples,.

He is still fully active in research on satellite communications, deep-space communications, radio propagation at millimetre waves. He has also researched literary texts according to Communication Theory and has proposed a mathematical theory on the deep-language structure of alphabetical texts, as well as identifying and studying multiple linguistic channels in literary texts and their connection with human short-term memory. According to Stanford University, he is in the top 2% of productive scientists in all fields.

He is in the Editorial Board of Information and Open Journal of Statistics and the Editor-in-Chief of Journal of Mathematical Studies of Literature. (Email: emilio.matricciani@polimi.it)