Revealing Short–Term Memory Communication Channels Embedded in Alphabetical Texts: Theory and Experiments

Emilio Matricciani *

Posted Date: 21 August 2025

doi: 10.20944/preprints202508.1588.v1

Keywords: Balto–Slavic languages; communication channels; language processing; Germanic languages; Greek; Latin; New Testament; Romance languages; short–term memory; translation; Uralic languages

 doi:10.20944/preprints202508.1588.v1

Article

Revealing Short–Term Memory Communication Channels Embedded in Alphabetical Texts: Theory and Experiments

Emilio Matricciani

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

Abstract

The aim of the present paper is to develop further a theory on the flow of linguistic variables making a sentence, namely, the transformation: (a) characters into words; (b) words into word intervals; (c) word intervals into sentences. The relationship between two linguistic variables is studied as a communication channel whose performance is determined by the slope of their regression line and by their correlation coefficient. The theory is applicable to any field/specialty in which a linear relationship holds between two variables. The signal–to–noise ratio Γ is a figure of merit of a channel being deterministic, i.e. a channel in which the scattering of the data around the regression line is negligible. The larger Γ is, the more the channel is deterministic. In conclusion, humans have invented codes whose sequences of symbols making words cannot vary very much for indicating single physical or mental objects of their experience (larger Γ). On the contrary, a large variability (smaller Γ) is achieved by introducing interpunctions to make word intervals, and word intervals to make sentences to communicate concepts.

Keywords: Balto–Slavic languages; communication channels; language processing; Germanic languages; Greek; Latin; New Testament; Romance languages; short–term memory; translation; Uralic languages

1. Introducing an Equivalent Input–Output Model of the Short–Term Memory

Humans can communicate and extract meaning both from spoken and written language. Whereas the sensory processing pathways for listening and reading are distinct, listeners and readers appear to extract very similar information about the meaning of a narrative story – heard or read – because the brain assimilates a written text like the corresponding spoken/heard text [1]. In the following, therefore, we consider the processing of reading or writing a text – a writer is also a reader of his/her own text – due to the same brain activity. In other words, the human brain represents semantic information in a modal form, independently of input modality.

How the human brain analyzes the parts of a sentence (parsing) and describes their syntactic roles is still a major question in cognitive neuroscience. In References [2,3], we proposed that a sentence is elaborated by the short–term memory (STM) with two independent processing units in series (equivalent surface processors), with similar size. The clues for conjecturing this input–output model emerged by considering many novels belonging to the Italian and English literatures. In Reference [3], we showed that there are no significant mathematical/statistical differences between the two literary corpora, according to the so–called surface deep–language parameters, suitably defined.

The model conjectures that the mathematical structure of alphabetical languages – digital codes created by the human mind for communication – seems to be deeply rooted in humans, independently of the particular language used or historical epoch. The complex and inaccessible

 doi:10.20944/preprints202508.1588.v1

mental process lying beneath communication – still largely unknown – can be studied by looking at the input–output functioning revealed by the structure of alphabetical languages.

The first processor is linked to the number of words between two contiguous interpunctions, variable indicated by I_p – termed word interval (Appendix A lists the mathematical symbols used in the present article) – approximately ranging in Miller's 7 ± 2 law range [4–30]. The second processor is linked to the number M_F of I_p's contained in a sentence, referred to as the extended short–term memory (E–STM), ranging approximately from 1 to 6. These two units can process sentences containing approximately a number of words from 8.3 to 61.2, values that can be converted into time by assuming a reading speed. This conversion gives $2.6 \sim 19.5$ seconds for a fast–reading reader [14], and $5.3 \sim 30.1$ seconds for a reader of novels, values well supported by experiments [15–30].

The E–STM must not be confused with the intermediate memory [31,32]. It is not modelled by studying neuronal activity, but by studying only the surface aspects of human communication due, of course, to neuronal activity, such as words and interpunctions, whose effects writers and readers experience since the invention of writing. In other words, the model proposed in References [2,3] describes the "input–output" characteristics of the STM. In Reference [33], we further developed the theory by including an equivalent first processor that memorizes syllables and characters to produce a word.

In conclusion, in References [2,3,33] we have proposed an input–output model of the STM, made of three equivalent linear processors in series, which independently process: (1) syllables and characters to make a word, (2) words and interpunctions to make a word interval; (3) word intervals to make a sentence. This is a simple, but a useful approach because the multiple brain process regarding speech/texts is not yet fully understood but characters, words and interpunctions – these latter are needed to distinguish word intervals and sentences – can be easily studied [34–36].

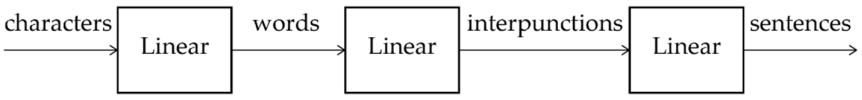
In other words, the model conjectures that the mathematical structure of alphabetical languages is deeply rooted in humans, independently of the particular language used or historical epoch. The complex and inaccessible mental process lying beneath communication – still largely unknown – is revealed by looking at the input–output functioning built–in in alphabetical languages of any historical epoch.

The literature on the STM and its various aspects is immense and multidisciplinary – we have recalled above only few references – but nobody – as far as we know – has considered the connections we found and discussed in References [2,3,33]. Our modelling of the STM processing by three units in series is new.

A sentence conveys meaning, of course, therefore the theory we have developed might be one of the necessary starting points to arrive at *the* Information Theory that will finally include meaning.

Today, many scholars are trying to arrive at a "semantic communication" theory or "semantic information" theory, but the results are still, in our opinion, in their infancy [37–45]. These theories, as those concerning the STM, have not considered the main "ingredients" of our theory, namely the number of characters per word C_P, I_P and M_F, parameters that anybody understands and can calculate in any alphabetical language [34–36], as a starting point for including meaning, still a very open issue.

Aim of the present paper is twofold: (a) to further develop the theory proposed in Reference [2,3,33], and (b) apply it to the flow of linguistic variables making a sentence. This "signal" flow is built–in in the model proposed in Reference [33], namely, the transformation of: (a) characters into words; (b) words into word intervals; (c) word intervals into sentences, according to Figure 1. Since the connection between these linguistic variables is described by regression lines [34–36], in the present article we analyze experimental scatterplots between these variables



 doi:10.20944/preprints202508.1588.v1

Figure 1. Flow chart of linguistic variables. The output variable of each block is connected to its input variable by a regression line.

The article is ideally divided in two parts. In the first part – from Section 2 to Section 4 – we recall and further develop the theory of linear channels [2,3,33]; in the second part – from Section 5 to Section 8 – we apply it to a significant database of literary texts.

The database of literary texts considered is a large set of the New Testament (NT) books, namely the Gospels according to *Matthew*, *Mark*, *Luke*, *John*, the Book of *Acts*, the *Epistle to the Romans*, and the *Apocalypse* – 155 chapters in total, according to the traditional subdivision of these texts. We have considered the original Greek texts and their translation to Latin and to 35 modern languages, texts partially studied in Reference [35]. Notice that in this paper, "translation" is indistinguishable from "language" because we deal only with one translation per language.

We consider the NT books and their modern translations for two reasons: (a) they tell the same story, therefore it is meaningful to compare the translations in different languages; (b) they use common words – not the words of scientific/academic disciplines – therefore, they can give some clues on how most humans communicate.

After this introductory section, Section 2 presents the theory of linear regression lines and associated communication channels; Section 3 presents the connection of single linear channels; Section 4 proposes and discusses the theory of series connection of single channels affected by noise; Section 5 reports an exploratory data analysis of the NT texts; Section 6 reports findings concerning single channels, Section 7 concerning series connection of channels and Section 8 concerning cross channels; finally Section 9 summarizes the main findings and indicates future studies.

2. Theory of Linear Regression Lines and Associated Communication Channels

In this section, we recall and further expand the general theory stochastic variables linearly connected, originally developed for linguistic channels [35,36] but applicable to any other field/specialty in which a linear relationship holds between two variables.

Let x (independent variable) and y (dependent variable) linked by the line:

$$y = mx + b \qquad (1)$$

Notice that Eq. (1) models a deterministic relationship through the slope m and the intercept b. Since in most scatterplots between linguistic variables $b = 0$, in the following we assume

$$b = 0 \qquad (2)$$

However, notice that if $b \neq 0$, the theory can be fully applied by defining a new dependent variable $\check{y} = y - b$.

In general, the relationship between x and y is not deterministic, i.e., given by Eq. (1), but stochastic (random). Eq. (1) models, in fact, two variables perfectly correlated – correlation coefficient $r = 1$ – characterized by a multiplicative "bias" m. In general, however, these conditions do not hold, therefore Eq. (1) can be written as:

$$y = mx + n \qquad (3)$$

In Eq. (3) n is an additive Gaussian stochastic variable with zero mean value [34–36], therefore Eq. (3) models a noisy linear channel. Notice that n must not be confused with an intercept b.

Figure 2 shows the flow chart describing Eq. (1) and Eq. (3) with a system/channel representation. The black box indicated with m represents the deterministic channel, i.e. Eq. (1); the black box indicated with r represents the parallel channel due to the scattering of y around the regression line. The additive noise n is a Gaussian stochastic variable with zero mean that makes the linear channel partially stochastic, namely "noisy".

 doi:10.20944/preprints202508.1588.v1

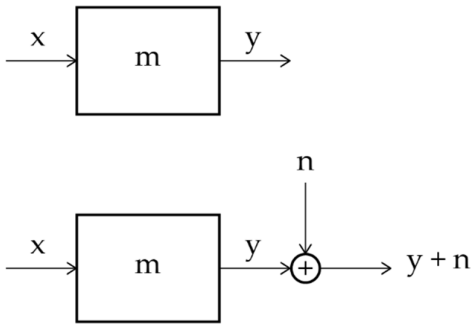


Figure 2. Flow chart in linear systems. Upper panel: deterministic channel with multiplicative bias m, Eq. (1). Lower panel: noisy deterministic channel with multiplicative bias and Gaussian noise source, Eq. (3).

Now, let us consider:

a) The variance of the difference between the values calculated with Eq. (1) ($m \neq 1$) and those calculated with $y = x$ ($m = 1$) (45° line) at given $x -$ value, as the "regression noise" power N_m [35]. This "noise" is due to the multiplicative bias between the two variables.

b) The variance of the difference between the values not lying on Eq. (1) ($r \neq 1$), and those lying on it ($r = 1$), as the "correlation noise" power N_r [35]. This "noise" is due to the spread of y around the line given by Eq. (1), modelled by n.

c) Let s_x^2 and s_y^2 be the variances of x and y.

In case (a), we get the difference $(m-1)x$, therefore the variance (or power) of the values lying on the regression line, regression noise, is given by:

$$N_m = (m-1)^2 s_x^2 \qquad (4)$$

Now, we define the regression noise–to–signal power ratio (NSR), R_m, as:

$$R_m = \frac{N_m}{s_x^2} = (m-1)^2 \qquad (5)$$

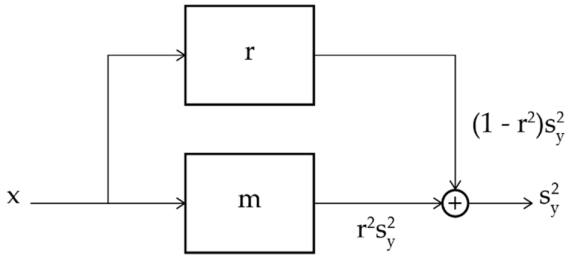
In case (b), the fraction of the variance s_y^2 due to the values of y not lying on the regression line (correlation noise power, N_r) is given by [46]:

$$N_r = (1-r^2)s_y^2 \qquad (6)$$

The parameter $\boldsymbol{r^2}$ is called the coefficient of determination and it is proportional to the variance of $\boldsymbol{y}$ explained by the regression line [46]. However, this variance is correlated with the slope $\boldsymbol{m}$ because the fraction of the variance $\boldsymbol{s_y^2}$ due to the regression line, namely $\boldsymbol{r^2 s_y^2}$, is related to $\boldsymbol{m}$ according to [46]:

$$r^2 s_y^2 = m^2 s_x^2 \qquad (7)$$

Figure 3 shows the flow chart of variances.



 doi:10.20944/preprints202508.1588.v1

Figure 3. Flow chart of variances: $r^2 s_y^2$ is the output variance of the values lying on the regression line, Eq. (7); $(1-r^2)s_y^2$ is the output variance due to the values of y not lying on the regression line, Eq. (6).

Therefore, inserting Eq. (7) in Eq. (6), we get the correlation NSR, R_r:

$$R_r = \frac{N_r}{s_x^2} = \frac{(1-r^2)}{r^2}\, m^2 \qquad (8)$$

Now, since the two noise sources are disjoint, the total NSR R of the channel shown in Figures 2, 3 is given by:

$$R = R_m + R_r \qquad (9)$$

Therefore, R depends only on the two parameters m and r of the regression line:

$$R = (m-1)^2 + \frac{(1-r^2)}{r^2}\, m^2 \qquad (10)$$

Finally, the signal–to–noise ratio (SNR) γ is given by:

$$\gamma = \frac{1}{R} = \frac{1}{(m-1)^2 + \frac{1-r^2}{r^2} m^2} \qquad (11)$$

In decibel:

$$\Gamma = 10 \times \log_{10}\gamma \qquad (12)$$

Of course, we expect that no channel can yield $|r| = 1$ and $|m| = 1$, therefore $\gamma = \infty$. In empirical scatterplots, very likely, $|r| < 1, |m| \neq 1$.

In conclusion, the slope m measures the multiplicative "bias" of the dependent variable y compared to the independent variable x in the deterministic channel; the correlation coefficient r measures how "precise" the linear best fit is.

Finally, notice the more direct and insightful analysis that can be achieved by using the NSR instead of the more common SNR because, in Eq. (9), the single channel NSRs simply add together. This makes easy to study, for example, which addend determines R, and thus Γ, while this is by far less easy with Eq. (11). Moreover, this choice leads also to a useful graphical representation of Eq. (10) that can guides analysis and design [11], as shown in Section 8.

In the next sections, we apply the theory of linear channel modeling to specific cases.

3. Connection of Single Linear Channels

We first study how the output variable y_k of channel k relates to the output variable y_j of another similar channel j for the same input x. This channel is termed "cross channel" and it is fundamental in studying language translation [35]. Secondly, we study how the output of a deterministic channel, modelled by Eq. (1), relates to the output of its stochastic version, Eq. (3).

3.1. Cross Channels

Let us consider a scatterplot k and a scatterplot j in which the independent variable x and the dependent variable y are linked by linear regression lines:

$$\boldsymbol{y_k = m_k x_k} \qquad (13)$$

$$\boldsymbol{y_j = m_j x_j} \qquad (14)$$

As discussed in Section 2, Eqs. (13)(14) do not give the full relationship between the two variables because they link only conditional average values, measured by the slopes $\boldsymbol{m_k}$ and $\boldsymbol{m_j}$ in the deterministic channels. According to Eq. (3), we can write more general linear relationships by considering the scattering of the data, always present in experiments, modelled by additive Gaussian zero–mean noise sources $\boldsymbol{n_k}$ and $\boldsymbol{n_j}$.

$$\boldsymbol{y_k = m_k x_k + n_k} \qquad (15)$$

$$\boldsymbol{y_j = m_j x_j + n_j} \qquad (16)$$

Now, we can develop a series of interesting investigations on these equations. By eliminating $\boldsymbol{x}$ we can compare the dependent variable $\boldsymbol{y_j}$ of Eq. (16) to the dependent variable $\boldsymbol{y_k}$ of Eq. (15) for

 doi:10.20944/preprints202508.1588.v1

$\boldsymbol{x_k = x_j = x}$. In doing so, we can find the regression line and the correlation coefficient of the new scatterplot linking $\boldsymbol{y_j}$ to $\boldsymbol{y_k}$ without the availability of the scatterplot itself.

By eliminating $\boldsymbol{x}$ between Eq.(15) and Eq. (16), we get:

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

Compared to the new independent variable $\boldsymbol{y_k}$, the slope $\boldsymbol{m_{kj}}$ of the regression line is given by:

$$\boldsymbol{m_{kj} = m_j / m_k} \quad (18)$$

Because the two Gaussian noise sources are independent and additive, the total noise is given by:

$$\boldsymbol{n_{kj} = -\frac{m_j}{m_k} n_k + n_j = -m_{kj} n_k + n_j} \quad (19)$$

Figure 4 shows the flow chart describing the cross–channel.

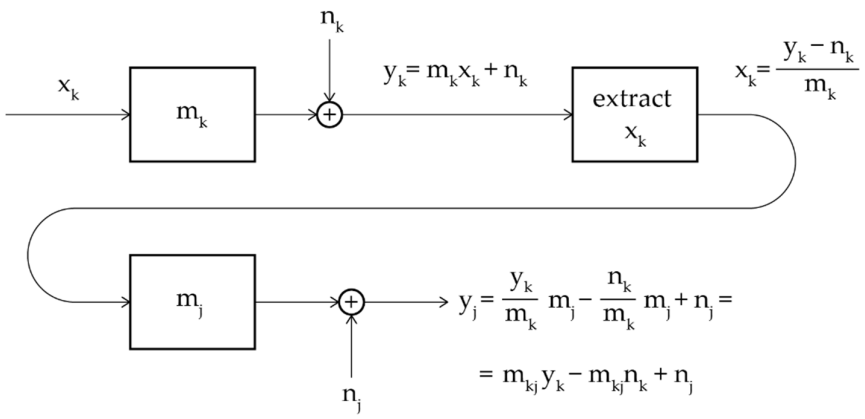


Figure 4. Flow chart describing the cross channel.

Now, from Eq. (18), $\boldsymbol{R_m}$ of the new channel is:

$$\boldsymbol{R_m = (m_{kj} - 1)^2} \quad (20)$$

The unknown correlation coefficient $\boldsymbol{r_{kj}}$ between $\boldsymbol{y_j}$ and $\boldsymbol{y_k}$ is given by [35]:

$$\boldsymbol{r_{kj} = \cos\left|\arccos(r_j) - \arccos(r_k)\right|} \quad (21)$$

Therefore, $\boldsymbol{R_r}$ of the new channel is:

$$\boldsymbol{R_r = \frac{1 - r_{kj}^2}{r_{kj}^2} m_{kj}^2} \quad (22)$$

In conclusion, in the new channel connecting y_j to y_k we can determine the slope and the correlation coefficient of the scatterplot between y_j and y_k for the same value of the independent variable x. Now, the availability of this scatterplot is experimentally very rare because it is unlikely to find values of y_k and y_j for exactly the same value of x, therefore cross channels can reveal relationships very difficult to discover experimentally.

In the next sections, we further developed the theory of linear channels, originally established in Reference [35] for cross channels.

3.2. Stochastic Versus Deterministic Channel

We compare a deterministic channel k with a stochastic channel j derived from channel k by adding noise. In other words, we start from the regression line given by Eq. (1) and then add noise n due to the correlation coefficient $|r| \neq 1$. Therefore, from the theory of stochastic channels discussed in Section 3.1, we get:

$$\boldsymbol{y_k = m_k x_k} \quad (23)$$

$$\boldsymbol{y_j = m_k x_k + n_k} \quad (24)$$

$$\boldsymbol{m_{kj} = m_k / m_k = 1} \quad (25)$$

$$\boldsymbol{R_m = (m_{kj} - 1)^2 = 0} \quad (26)$$

$$\boldsymbol{r_{kj} = \cos\left|\arccos(r_j) - \arccos(1)\right| = \cos\left|\arccos(r_j)\right| = r_j = r} \quad (27)$$

 doi:10.20944/preprints202508.1588.v1

$$\boldsymbol{R = R_r = \frac{1-r^2}{r^2}} \quad (28)$$

In conclusion, in transforming a deterministic channel into a stochastic channel only the correlation noise is present, therefore the SNR is given by:

$$\boldsymbol{\gamma = \frac{r^2}{1-r^2}} \quad (29)$$

Eq. (29) coincides with the ratio between the variance explained by the regression line (proportional to the coefficient of determination $\boldsymbol{r^2}$) and the variance due to the scattering (correlation noise), proportional to $\boldsymbol{1 - r^2}$ [46].

So far, we have considered single channels. In the next section, we consider the series connection of single channels to determine the SNR of the overall channel.

4. Series Connection of Single Channels Affected by Correlation Noise

In this section, we consider a channel made of series of single channels. We consider this case because it can be found in many specialties, and because in Section 7 we apply it to specific linguistic channels.

Figure 5 shows the flow chart of three single channels in series. These channels can be characterized as done in Section 3.2, i.e., only with the correlation noise, therefore, the overall channel is compared to the deterministic channel in which:

$$m = m_1 m_2 m_3 \quad (30)$$

From Figure 5, it is evident that the output noise of a preceding channel produces additive noise at the output of the next channel in series. The purpose of this section is to calculate R at the output of the series of channels.

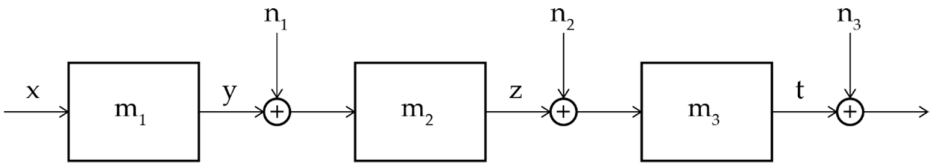


Figure 5. Flow chart of noisy single channels connected in series.

Theorem. The NSR R of n linear channels in series, each characterized by the correlation noise–to–signal ratio R_i, is given by:

$$R = \sum_{i=1}^{n} R_i \quad (30)$$

Proof. Let the three linear relationships of the isolated channels of Figure 5 (i.e., before connecting them in series) given by:

$$y = m_1 x + n_1 \quad (31)$$

$$z = m_2 y + n_2 \quad (32)$$

$$t = m_3 z + n_3 \quad (33)$$

Let s_x^2, s_y^2, s_z^2 the variances (power) of the variables, and let $N_1 = N_{1r}$, $N_2 = N_{2r}$, $N_3 = N_{3r}$ the variances (power) of the Gaussian zero–mean noise n_1, n_2, n_3, then the NSRs of the isolated channels are given by:

$$R_1 = \frac{N_1}{s_y^2} = \frac{N_1}{m_1^2 s_x^2} \quad (34)$$

$$R_2 = \frac{N_2}{s_z^2} = \frac{N_2}{m_2^2 s_y^2} \quad (35)$$

$$R_3 = \frac{N_3}{s_t^2} = \frac{N_3}{m_3^2 s_z^2} \quad (36)$$

When the first two blocks are connected in series, the input to the second block must include also the output noise of the first block, therefore from Eqs. (31)(33) we get the modified output variable $\check{z}$:

$$\check{z} = m_2 y = m_2(m_1 x + n_1) + n_2 = m_2 m_1 x + m_2 n_1 + n_2 \quad (37)$$

 doi:10.20944/preprints202508.1588.v1

In eq. (37), $m_2 m_1 x$ is the output "signal" and $m_2 n_1 + n_2$ is the output noise, therefore, the NSR at the output of the second block is:

$$R = \frac{m_2^2 N_1 + N_2}{m_2^2 m_1^2 s_x^2} = \frac{m_2^2 N_1}{m_2^2 m_1^2 s_x^2} + \frac{N_2}{m_2^2 m_1^2 s_x^2} = \frac{N_1}{m_1^2 s_x^2} + \frac{N_2}{m_2^2 s_x^2} = R_1 + R_2 \quad (38)$$

Now, for three channels in series, it is sufficient to consider R given by Eq. (38) as the input NSR to the third single channel to obtain the final NSR and prove Eq.(30):

$$R = R_1 + R_2 + R_3 \quad (39)$$

Finally, notice that R of Eq. (30) is proportional to the mean $< R_i >$:

$$R = n \sum_{i=1}^{n} R_i / n = n < R_i > \quad (40)$$

In other words, the series channel averages the single R_i.

In conclusion, Eqs. (30) – (40) allow studying channels made by the series of several single channels affected by correlation noise by simply adding their single NSRs.

In the next sections, we apply the theory to linguistic channels suitably defined, after exploring the database on the NT mentioned in Section 1.

5. Exploratory Data Analysis

In this second part, we explore the linear relationships between characters, words, interpunctions and sentences, according to the flow chart shown in Figure xx, of the New Testament books considered (*Matthew*, *Mark*, *Luke*, *John*, *Acts*, *Epistle to the Romans*, *Apocalypse*). This is the database of our experimental analysis and application of the theory of linear channels discussed in the previous sections.

Table 1 lists language of translation and language family, with total number of characters (C), words (W), sentences (S) and interpunctions (I).

Table 1. Language of translation and language family of the New Testament books (*Matthew*, *Mark*, *Luke*, *John*, *Acts*, *Epistle to the Romans*, *Apocalypse*), with total number of characters (C), words (W), sentences (S) and interpunctions (I). The list concerning the genealogy of Jesus of Nazareth reported in *Matthew* 1.1–1.17 17 and in *Luke* 3.23–3.38 was deleted for not biasing the statistics of linguistic variables [35]. The source of the texts considered is reported in Reference [35].

Language	Order	Abbreviation	Language Family	C	W	S	I
Greek	1	Gr	Hellenic	486520	100145	4759	13698
Latin	2	Lt	Italic	467025	90799	5370	18380
Esperanto	3	Es	Constructed	492603	111259	5483	22552
French	4	Fr	Romance	557764	133050	7258	17904
Italian	5	It	Romance	505535	112943	6396	18284
Portuguese	6	Pt	Romance	486005	109468	7080	20105
Romanian	7	Rm	Romance	513876	118744	7021	18587
Spanish	8	Sp	Romance	505610	117537	6518	18410

 doi:10.20944/preprints202508.1588.v1

Danish	9	Dn	Germanic	541675	131021	8762	22196
English	10	En	Germanic	519043	122641	6590	16666
Finnish	11	Fn	Germanic	563650	95879	5893	19725
German	12	Ge	Germanic	547982	117269	7069	20233
Icelandic	13	Ic	Germanic	472441	109170	7193	19577
Norwegian	14	Nr	Germanic	572863	140844	9302	18370
Swedish	15	Sw	Germanic	501352	118833	7668	15139
Bulgarian	16	Bg	Balto–Slavic	490381	111444	7727	20093
Czech	17	Cz	Balto–Slavic	416447	92533	7514	19465
Croatian	18	Cr	Balto–Slavic	425905	97336	6750	17698
Polish	19	Pl	Balto–Slavic	506663	99592	8181	21560
Russian	20	Rs	Balto–Slavic	431913	92736	5594	22083
Serbian	21	Sr	Balto–Slavic	441998	104585	7532	18251
Slovak	22	Sl	Balto–Slavic	465280	100151	8023	19690
Ukrainian	23	Uk	Balto–Slavic	488845	107047	8043	22761
Estonian	24	Et	Uralic	495382	101657	6310	19029
Hungarian	25	Hn	Uralic	508776	95837	5971	22970
Albanian	26	Al	Albanian	502514	123625	5807	19352
Armenian	27	Ar	Armenian	472196	100604	6595	18086
Welsh	28	Wl	Celtic	527008	130698	5676	22585
Basque	29	Bs	Isolate	588762	94898	5591	19312
Hebrew	30	Hb	Semitic	372031	88478	7597	15806

 doi:10.20944/preprints202508.1588.v1

Cebuano	31	Cb	Austronesian	681407	146481	9221	16788
Tagalog	32	Tg	Austronesian	618714	128209	7944	16405
Chichewa	33	Ch	Niger–Congo	575454	94817	7560	15817
Luganda	34	Lg	Niger–Congo	570738	91819	7073	16401
Somali	35	Sm	Afro–Asiatic	584135	109686	6127	17765
Haitian	36	Ht	French Creole	514579	152823	1042 9	23813
Nahuatl	37	Nh	Uto–Aztecan	816108	121600	9263	19271

Figure 6 shows the scatterplots in the original Greek texts between: (a) characters and words; (b) words and interpunctions; (c) interpunctions and sentences; (d) characters and sentences. Figure 7 shows these scatterplots in the English translation. Appendix B shows example of scatterplots in other languages. Table 2 reports slope m and correlation coefficient r of the indicated scatterplots (155 samples for each scatterplot) for each translation, namely the input parameters of our theory on communication channels. The differences between languages are due to the large "domestication" of the original Greek texts discussed in Reference [47].

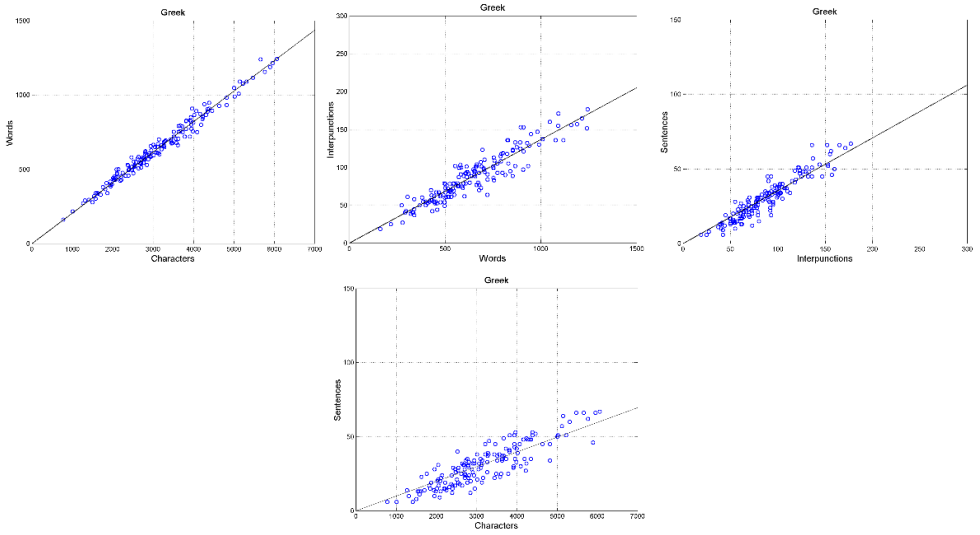
The four scatterplots define fundamental linear channels and they are connected with important linguistic parameters, previously studied [34–36], namely:

a) The number of characters per word, C_P, given by the ratio between characters (abscissa) and words (ordinate) in Figure 6(a).
b) The number of words between two successive interpunctions, I_P – called the word interval – given by the ratio between interpunctions (abscissa) and words (ordinate) in Figure 6(b).
c) The number of word intervals in sentences, M_F, given by the ratio between sentences (abscissa) and interpunctions (interpunctions) in Figure 6(c).

Figure 6(d) shows the scatterplot between characters and sentences, which will be discussed in Section 7.

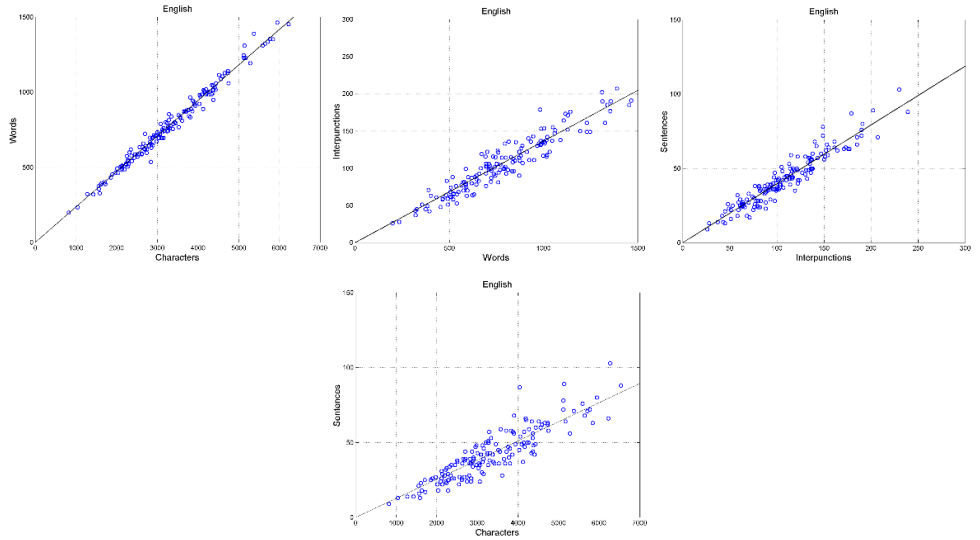
In the next section, we study the cannels corresponding to these scatterplots.

 doi:10.20944/preprints202508.1588.v1



(**a**) (**b**) (c) (d)

Figure 6. Scatterplots in the original Greek texts between: (a) characters and words; (b) words and interpunctions; (c) interpunctions and sentences; (d) between characters and sentences.



(**a**) (**b**) (c) (d)

Figure 7. Scatterplots in the English texts between: (a) characters and words; (b) words and interpunctions; (c) interpunctions and sentences; (d) between characters and sentences. In this case, English is the language to be translated.

Table 2. Slope m and correlation coefficient r, of the indicated regression lines in each language/translation.

Language	Words vs Characters		Interpunctions vs Words		Sentences vs Interpunctions		Sentences vs Characters.	
	m_1	r_1	m_2	r_2	m_3	r_3	m	r
Greek	0.2054	0.9893	0.1369	0.9298	0.3541	0.9382	0.0099	0.8733
Latin	0.1944	0.9890	0.2038	0.9515	0.2957	0.9366	0.0117	0.8646
Esperanto	0.2256	0.9920	0.2045	0.9668	0.2461	0.9545	0.0113	0.8998
French	0.2386	0.9945	0.1347	0.9483	0.4045	0.9509	0.0131	0.9339
Italian	0.2233	0.9921	0.1636	0.9476	0.3489	0.9537	0.0127	0.8856

 doi:10.20944/preprints202508.1588.v1

Portuguese	0.2246	0.9924	0.1845	0.9620	0.3532	0.9484	0.0146	0.9106
Romanian	0.2312	0.9933	0.1568	0.9589	0.3823	0.9384	0.0138	0.8820
Spanish	0.2320	0.9919	0.1580	0.9619	0.3565	0.9581	0.0130	0.9047
Danish	0.2417	0.9945	0.1694	0.9574	0.3961	0.9551	0.0163	0.9257
English	0.2364	0.9925	0.1365	0.9509	0.3962	0.9483	0.0128	0.8916
Finnish	0.1702	0.9904	0.2067	0.9621	0.3029	0.9464	0.0107	0.9131
German	0.2142	0.9938	0.1731	0.9637	0.3511	0.9555	0.0130	0.9325
Icelandic	0.2315	0.9937	0.1805	0.9600	0.3672	0.9527	0.0154	0.9296
Norwegian	0.2460	0.9956	0.1305	0.9581	0.5018	0.9621	0.0162	0.9626
Swedish	0.2371	0.9918	0.1277	0.9218	0.5041	0.9499	0.0154	0.9423
Bulgarian	0.2271	0.9926	0.1809	0.9590	0.3861	0.9482	0.0159	0.9203
Czech	0.2223	0.9927	0.2125	0.9496	0.3879	0.9282	0.0184	0.9034
Croatian	0.2287	0.9915	0.1825	0.9504	0.3853	0.9605	0.0161	0.9095
Polish	0.1968	0.9939	0.2159	0.9650	0.3768	0.9245	0.0160	0.9049
Russian	0.2148	0.9889	0.2397	0.9712	0.2566	0.9274	0.0132	0.8728
Serbian	0.2370	0.9925	0.1745	0.9513	0.4154	0.9436	0.0172	0.9111
Slovak	0.2149	0.9911	0.1973	0.9532	0.4085	0.9544	0.0173	0.9092
Ukrainian	0.2181	0.9893	0.2122	0.9730	0.3556	0.9448	0.0166	0.9545
Estonian	0.2054	0.9912	0.1881	0.9559	0.3342	0.9467	0.0129	0.8995
Hungarian	0.1882	0.9885	0.2412	0.9719	0.2632	0.9482	0.0120	0.9282
Albanian	0.2458	0.9896	0.1573	0.9607	0.3040	0.9582	0.0117	0.9106
Armenian	0.2140	0.9753	0.1802	0.9699	0.3698	0.9635	0.0142	0.8868
Welsh	0.2482	0.9953	0.1734	0.9818	0.2543	0.9493	0.0109	0.9336
Basque	0.1614	0.9939	0.2045	0.9673	0.2925	0.9506	0.0097	0.9210
Hebrew	0.2380	0.9945	0.1784	0.9615	0.4869	0.9635	0.0206	0.9144
Cebuano	0.2149	0.9983	0.1145	0.9465	0.5491	0.9578	0.0136	0.9670
Tagalog	0.2072	0.9957	0.1281	0.9555	0.4879	0.9363	0.0130	0.9411
Chichewa	0.1649	0.9964	0.1685	0.9420	0.4733	0.9596	0.0132	0.9381
Luganda	0.1610	0.9951	0.1797	0.9488	0.4314	0.9501	0.0125	0.9235
Somali	0.1876	0.9965	0.1628	0.9300	0.3505	0.9399	0.0107	0.8773
Haitian	0.2972	0.9959	0.1571	0.9672	0.4338	0.9567	0.0203	0.9288
Nahuatl	0.1489	0.9955	0.1593	0.9304	0.4759	0.9582	0.0114	0.9435
Overall	0.2161 ±0.0296	0.9925 ±0.0038	0.1750 ±0.0308	0.9558 ±0.0131	0.3795 ±0.0755	0.9492 ±0.0100	0.0140 ±0.0027	0.9149 ±0.0252

Figure 8 shows the probability distributions of the correlation coefficient r and the coefficient of determination r^2, for the scatterplots: words versus characters (green line); interpunctions versus words (cyan); sentences versus interpunctions (magenta). The black line refers to the scatterplot sentences versus characters; the red line refers to the series channel considered in Section 7, which links characters to sentences.

On correlation coefficients – and consequently, on the coefficient of determination, which determines the SNR – we notice the following remarkable findings:

 doi:10.20944/preprints202508.1588.v1

a) In any language, the largest correlation coefficient is found in the scatterplot between characters and words. The communication digital codes invented by humans show remarkable strict relationships between digital symbols (characters) and their sequences (words) to indicate items of their experience, material or immaterial. Languages do not differ from each other very much being r in the range $0.9753 - 0.9983$ (Armenian, Cebuano) and being, overall $r = 0.9925 \pm 0.0038$.

b) The smallest correlation coefficient is found in the scatterplot between characters and sentences, overall 0.0140 ± 0.0027. This relationship must be, of course, the most unpredictable and variable because the many digital symbols that make a sentence can create an extremely large number of combinations, each delivering a different concept.

c) The correlation coefficient (and also the coefficient of determination r^2) decreases as characters combine to create words, as words combine to create word intervals and as word intervals combine to create sentences.

The path just mentioned in item c) describes an increasing creativity and variety of meaning, other than that of the deterministic channel.

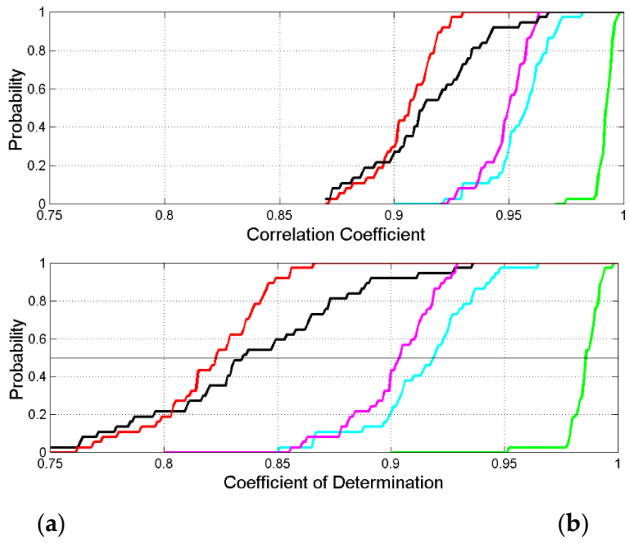


Figure 8. (a) Probability distribution of the correlation coefficient r; **(b)** probability distribution of the coefficient of determination r^2. Both refer to the following scatterplots: words versus characters, green; interpunctions versus words: cyan; sentences versus interpunctions, magenta. The black line refers to the scatterplot sentences versus characters; the red line refers to the series channel considered in Section 7.

The characters–to–words channel shows the smallest r^2, therefore, this channel is the nearest to be purely deterministic. It does not tend to be typical of a particular text/writer but more of a language because a writer has very little freedom in using words of very different length [34], if we exclude specialized words belonging to scientific and academic disciplines.

On the contrary, the channels words–to–interpunctions and the interpunctions–to–sentences are less deterministic, a writer can exercise his/her creativity of expression more freely, therefore these channels depend more on writer/text than on language. Finally, the big "jump" from characters to sentences gives the greatest freedom.

In conclusion, humans have invented codes whose sequences of symbols making words cannot variate very much for indicating single physical or mental objects of their experience. To communicate concepts, on the contrary, a large variability can be achieved by introducing interpunctions to form word intervals and word intervals to form sentences, the final depositary of human basic concepts.

 doi:10.20944/preprints202508.1588.v1

Figure 9 shows the probability distributions of the slope m. The black line (only partially visible because it is superposed to the red line) refers to the scatterplot sentences versus characters; the red line refers to the series channel that connects sentences to characters, discussed in Section 7.

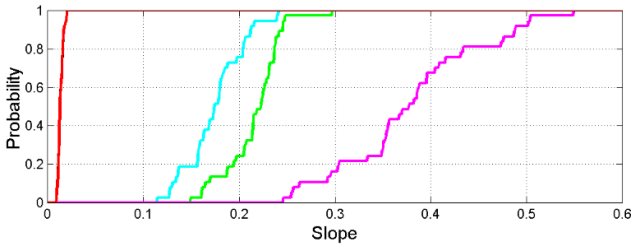


Figure 9. Probability distribution of the regression line slope m in the following scatterplots: words versus characters, green; interpunctions versus words: cyan; sentences versus interpunctions, magenta. The black line (not visible because superposed by the red cline) refers to the scatterplot sentences versus characters; the red line refers to the series channel considered in Section 7.

On slopes, we notice the following important findings:

a) The slope of the scatterplot between interpunctions and sentences (magenta line) is the largest in any language– overall 0.3795 ± 0.0755 – and determines the number of word intervals, M_F, contained in a sentence in its deterministic channel.

b) The slope of the scatterplot between interpunctions and words (cyan line) determines the length of the word interval, I_P, in its deterministic channel.

c) The slope of the scatterplot between words and characters (green line) determines the number of characters per word, C_P, in its deterministic channel. As discussed below, this channel is the most "universal" channel because, from language to language C_P varies little, compared to other linguistic variables.

d) The smallest slopes are in the scatterplots between characters and sentences, overall 0.0140 ± 0.0027. For example, in English there are 519,043 characters and 6590 sentences (Table 1); now, according to Table 1, the deterministic channel predicts $0.0128 \times 519{,}043 \approx 6644$ sentences just $+0.8$ % difference from the true value.

As reiterated above, the slopes describe deterministic channels. As discussed in Section 6, a deterministic channel is not "deterministic" for what concerns the number of concepts, because the same number of sentences can communicate different meanings by just changing words and interpunctions. What is "deterministic" is the size of the ensemble.

In the next section, we model single linguistic channels, i.e. channels not yet connected in series from the linear relationships shown above.

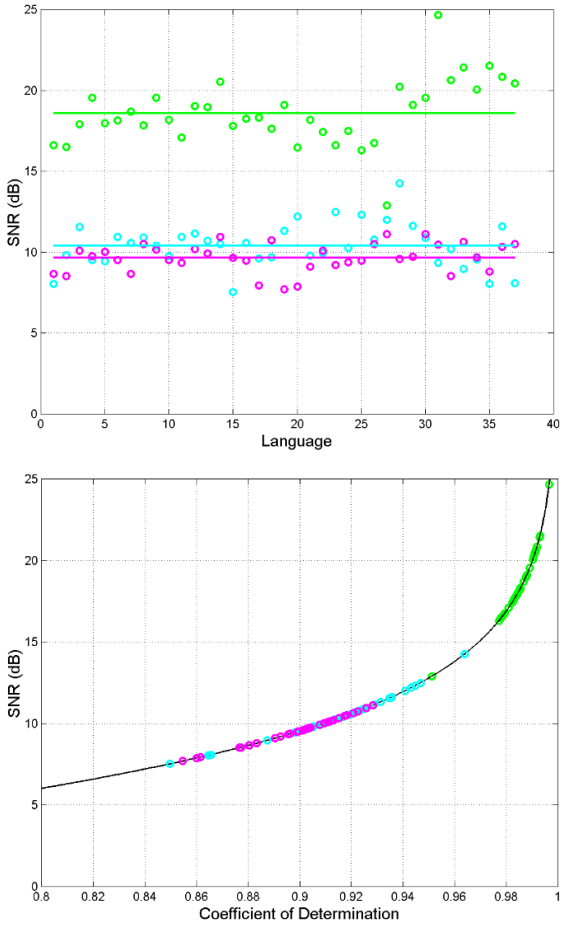
6. Single Linguistic Channels

In this section, we apply the theory developed in Section 3.2 to the scatterplots of Section 5, therefore, to the following single channels:

(a) Characters–to–Words.
(b) Words–to–Interpunctions.
(c) Interpunctions–to–Sentences.

These single channels are modelled like in Figures 2, 3, they are affected only by the regression noise. $\boldsymbol{\Gamma}$ is obtained from Eq. (29) and drawn in Figure 10. Table 3 reports mean and standard deviation of $\boldsymbol{\Gamma}$ in each channel.

 doi:10.20944/preprints202508.1588.v1



(**a**) (**b**)

Figure 10. (a) Signal–to–noise ratio SNR Γ (dB) versus language (see order number in Table 1); (b) theoretical relationship between Γ and the coefficient of determination. Characters–to–words, green; words–to–interpunctions, cyan; interpunctions–to–sentences, magenta. The horizontal lines in (a) draw mean values.

Table 3. Mean and standard deviation of the signal–to–noise ratio Γ (dB) in the indicated channel. The probability density function of each channel is modelled as Gaussian.

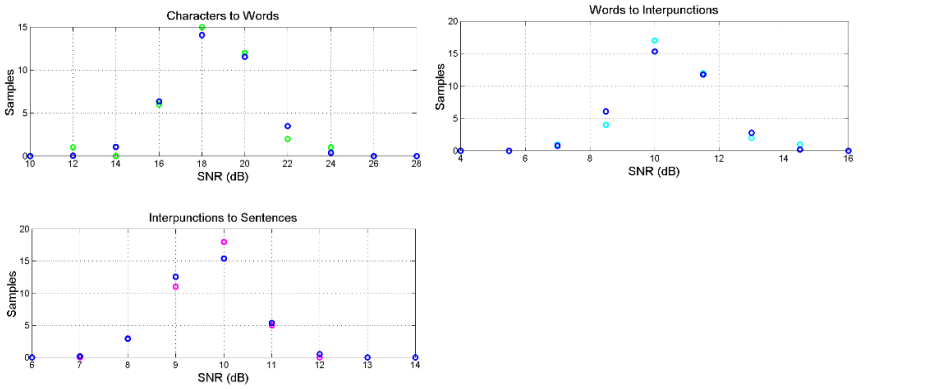
Channel	**Mean ± Standard deviation of Γ (dB)**
Characters–to–Words	18.60 ± 2.00
Words–to–Interpunctions	10.42 ± 1.38
Interpunctions–to–Sentences	9.66 ± 0.89

From Figure 10 and Table 3, we notice the following interesting facts:

a) Languages show different **Γ** due to the large degree of domestication of the original Greek texts [47].
b) **Γ** decreases steadily in this order: characters–to–words, words–to–interpunctions, interpunctions–to–sentences. A decreasing **Γ** says, relatively, how much a channel is less deterministic.
c) Words–to–interpunction and interpunctions–to–sentences have close values, therefore they show similar deterministic channels.
d) Most languages have **Γ** greater than that in Greek. This agrees with the finding that in modern translations of the Greek texts domestication prevails over foreignization [47].
e) Finally, we can consider **Γ** as a figure of merit of a linguistic channel being deterministic: the larger **Γ** is, the more the channel is deterministic.

 doi:10.20944/preprints202508.1588.v1

Figures 11 shows histograms (37 samples) of $\boldsymbol{\Gamma}$ for each channel. The probability density function of $\boldsymbol{\Gamma}$ can be modelled with a Gaussian model (therefore, $\boldsymbol{\gamma}$ is a lognormal stochastic variable) with mean and standard deviation reported in Table 3. Figure 12 shows the probability functions of $\boldsymbol{\Gamma}$ which show, again, differences and similarities of the channels.



(a) **(b)** **(c)**

Figure 11. Histograms (37 samples) of the signal–to–noise ratio SNR Γ for each channel: **(a)** characters–to–interpunction; **(b)** words–to–interpunctions; **(c)** interpunctions–to–sentences**.**

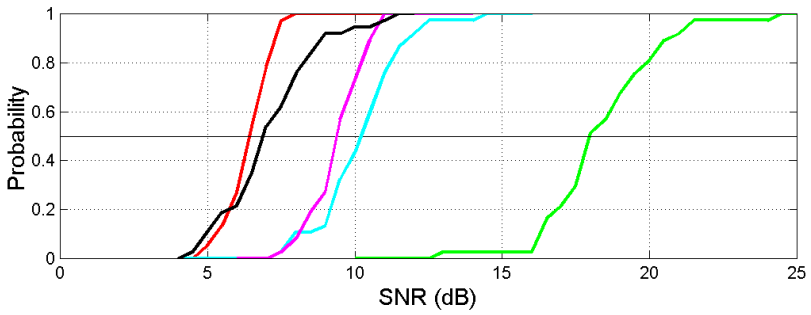


Figure 12. Probability distribution of the the signal–to–noise ratio SNR Γ in the following channels: characters–to–words, green; words–to–interpunctions, cyan; interpunctions–to–sentences, magenta. The black line refers to the channel characters–to–sentences estimated form the scatterplot of Figure 6(d); the red line refers to the series channel considered in Section 7.

In conclusion, the large $\boldsymbol{\Gamma}$ of the characters–to–words channel, in any language, indicates that the transformation of characters into words is the most deterministic.

In the next Section, we connect the single channels to obtain the series channels modelled in Figure 5 and study them according to the theory of Section 4.

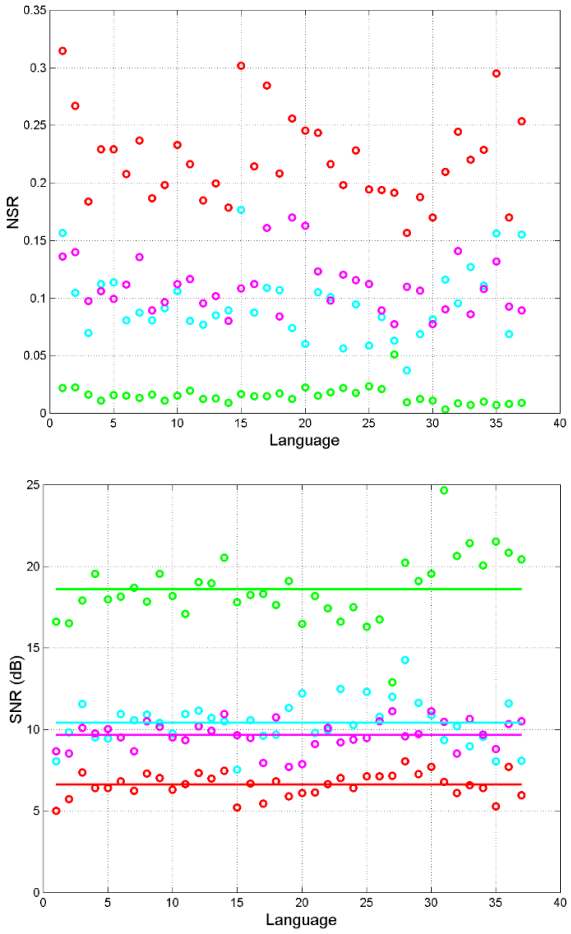
7. Series Connection of Linguistic Channels Affected by Correlation Noise

Let us connect the three single channels to obtain the series channel shown in Figure 5 and apply the theory of Section 4. We first show the results concerning the theory of series channel, and then we compare the single channel characters–to–sentences to that obtained with the series of single channels.

Figure 13(a) shows the single NSRs and the series NSRs in linear units for each language; Figure 13(b) shows the corresponding $\boldsymbol{\Gamma}$ (dB), partially already reported in Figure 10. We can notice that in the sum indicated in Eq.(30), the NSR of the characters–to–words channel is negligible compared to the other two NSRs. For example, in English (language no. 10)$\boldsymbol{R = 0.0152 + 0.1059 + 0.1120 =}$

 doi:10.20944/preprints202508.1588.v1

$\mathbf{0.2331} \approx \mathbf{0.1059} + \mathbf{0.1120} = \mathbf{0.2179}$, therefore $\boldsymbol{R} \approx \mathbf{0.22}$ against $\boldsymbol{R} \approx \mathbf{0.23}$. In general, $\boldsymbol{R_1} \ll \boldsymbol{R_2}, \boldsymbol{R_3}$ so that the characters–to–words channel can be ignored, to a first approximation, because it is about $\mathbf{1/10}$ of the other two addends in Eq. (30).

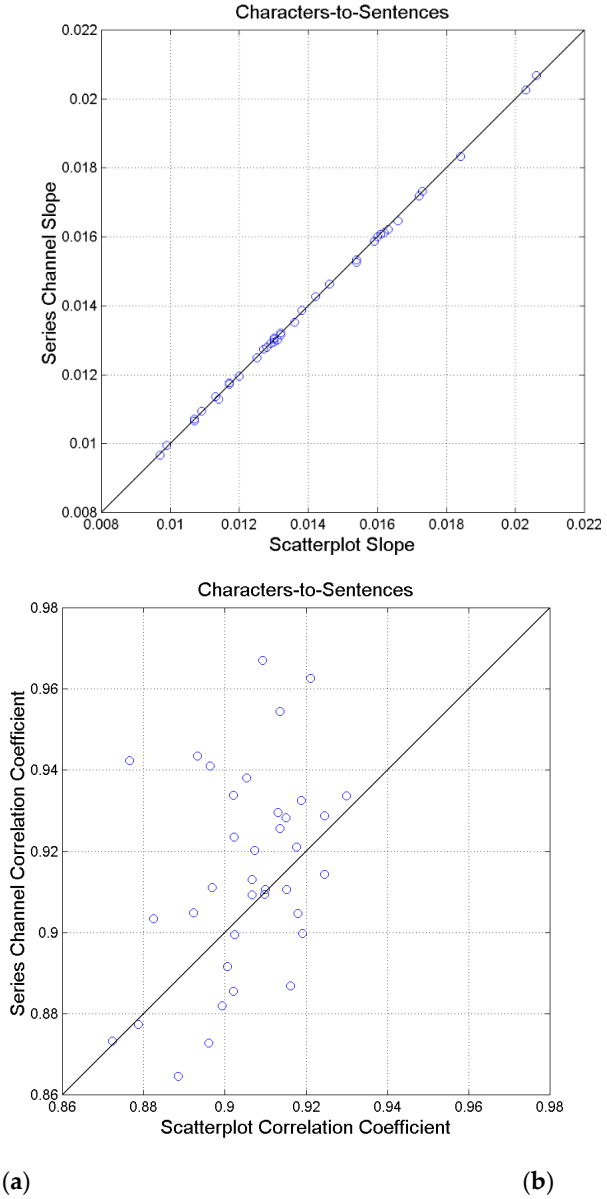


(**a**) (**b**)

Figure 13. (a) Single channel NSR and series channel NSR in linear units; **(b)** signal–to–noise ratio SNR Γ (dB). The horizontal lines draw mean values. Channels: characters–to–words, green; words–to–interpunctions, cyan; interpunctions–to–sentences, magenta; series channel, red.

For the characters–to–words channel, Figure 14(a) shows the slope calculated from the scatterplot between characters and sentences (Table 2) and the slope given by Eq. (30). The agreement is excellent, in practice the two values coincide (correlation coefficient 0.9998). Figure 14(b) shows scatterplot between the correlation coefficient calculated from the scatterplot between characters and sentences (Table 2) and that calculated by solving Eq. (29) for $\boldsymbol{r}$ after calculating $\boldsymbol{\gamma}$ from Eq. (39). In this case, the two values are poorly correlated (correlation coefficient 0.3929). Finally, notice the difference between the probability of $\boldsymbol{\Gamma}$ calculated by solving Eq. (29) for $\boldsymbol{r}$ – red line in Figure 12 – and that calculated from the available scatterplots and regression line (Table 2), black line. The smoother red curve models more accurately the relationship between characters and sentences than the available scatterplot shown in Figure 6(d), because $\boldsymbol{R}$ is proportional to the mean value of the single channels, see Eq. (40).

 doi:10.20944/preprints202508.1588.v1



(**a**) (**b**)

Figure 14. (a) Scatterplot between the slope calculated from the scatterplot between characters and sentences (Table xx) and the slope given by Eq. (30); (b) Scatterplot between the correlation coefficient calculated from the scatterplot between characters and sentences (Table 2) and that calculated by solving Eq. (29) for the r.

In conclusion, $\mathbf{\Gamma}$ calculated in a series channel linking two variables is more reliable than that calculated from a single channel/scatterplot between the two variables.

In the next section, we apply the theory of cross–channels of Section 3.1.

8. Cross Channels: Language Translations

In cross channels we study how the output variable y_k of channel k relates to the output variable y_j of another similar channel j for the same input x, therefore we apply the theory of Section 3.1. In this new channel, we can determine the slope and the correlation coefficient of the scatterplot between y_k and y_j for the same value of the independent variable x, therefore cross channels can reveal relationships more difficult to discover experimentally.

From the database of the NT texts and the scatterplots of Figure 6, we can study at least three cross channels:

 doi:10.20944/preprints202508.1588.v1

a) The words–to–words channel, by eliminating characters, therefore the number of words are compared for the same number of characters.

b) The interpunctions–to–interpunctions channel, by eliminating words, therefore, the number of words intervals are compared for the same number of words.

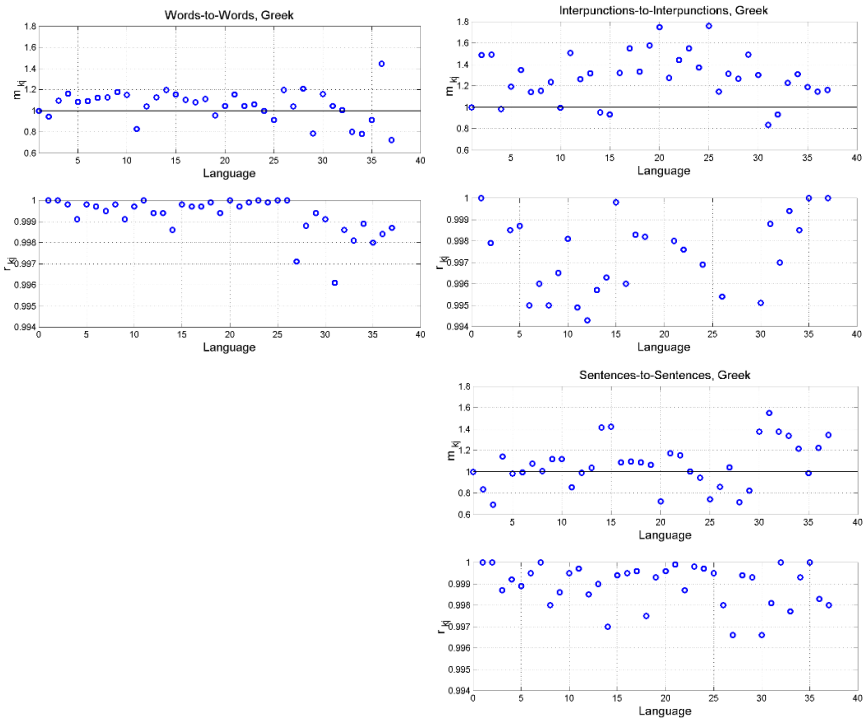
c) The sentences–to–sentences channel, by eliminating interpunctions, therefore, the number of sentences are compared for the same number of word intervals.

Now, since these channels connect one independent variable in one language to the same (dependent) variable in another language, they describe very important linguistic channels, namely translation channels and they can be studied from this particular perspective. Therefore, cross channels in alphabetical texts describe the mathematics/statistics of translation, as we first studied in Reference [35].

Figure 15 shows the slope m_{kj} and the correlation coefficient r_{kj} by assuming Greek as language k, namely the reference language, for the three cross channels We can notice the following:

a) For most languages, $m_{kj} > 1$ in any cross channel, therefore most modern languages tend to use more words – for the same number of characters –; more word intervals –for the same number of words – and more sentences – for the same number of word intervals – than Greek. In other words, the corresponding deterministic channel (the channel characterized a multiplicative slope) is significantly biased compared to the original Greek texts.

b) The correlation coefficient r_{kj} is always very near unity. Therefore the scattering of the data around the regression line is similar in all the three cross channels.



(**a**) (**b**) (**c**)

Figure 15. Mean value (upper panel) and correlation coefficient (lower panel) in the indicated languages, assuming Greek as reference language (to be translated) in the channels: (a) words–to–words; (b) interpunctions–to–interpunctions; (c) sentences–to–sentences.

Figure 16 shows the findings assuming English as reference language. In this case, we consider the "translation" from English into the other languages [35]. Clear differences are noticeable:

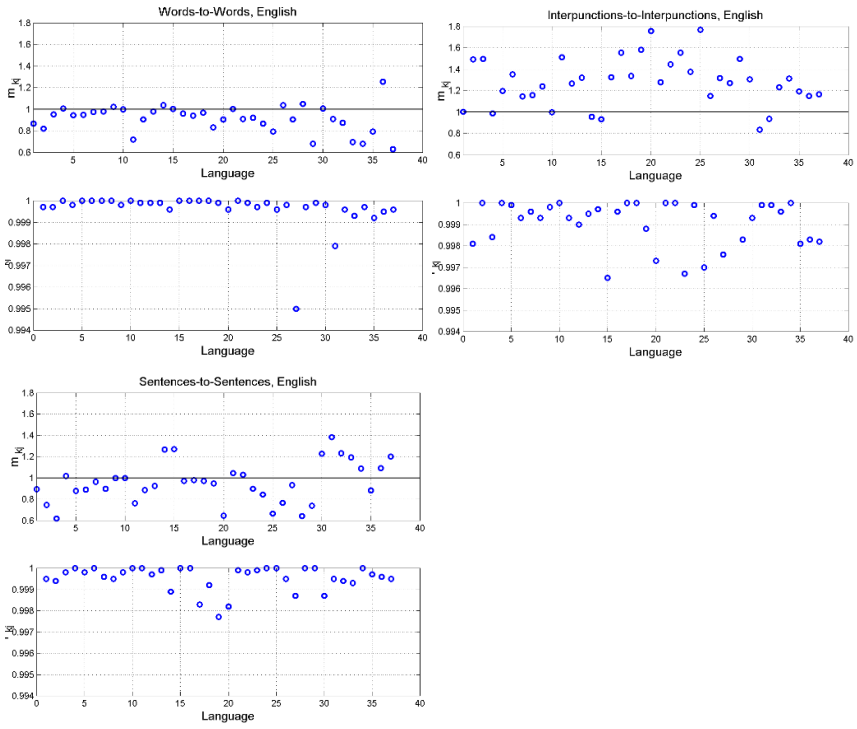
 doi:10.20944/preprints202508.1588.v1

a) Words–to–words channel: for most languages $m_{kj} \leq 1$. The multiplicative bias is small, language tend to use the same number of words of English. This was not the case for Greek. The correlation coefficient r_{kj} is practically unity for all languages. In other words, modern language tend to use the same number of words English, for the same number of characters, therefore domestication of the alleged translation of English to the other languages is moderate, compared to Greek or Latin (see languages 1 and 2 in Figure 16(a)).

b) Interpunctions–to–interpunctions channel: the multiplicative bias m_{kj} is strong, as in is Greek, therefore, the deterministic cross channels are different from language to language. The correlation coefficient r_{kj} is more scattered than in Figure 15, and different from language to language. Curiously, in the channel English–to–Greek, $m_{kj} \approx 1$, no bias The correlation coefficient r_{kj} is similar to that of the sentences–to–sentences channel.

c) Sentences–to–sentences: $m_{kj} \leq 1$ for most languages, r_{kj} is similar to that of the interpunctions–to–interpunctions channel.

Since similar diagrams can be shown when other modern languages are considered as the independent language – not shown for brevity – we can conclude that the translation from Greek to modern languages show a high degree of domestication, due especially to the multiplicative bias, namely to the deterministic channels not to the stochastic part of the channel. In conclusion, the translation from a modern language into another modern language is mainly done through deterministic channels. Therefore, the SNR is mainly determined by R_m.

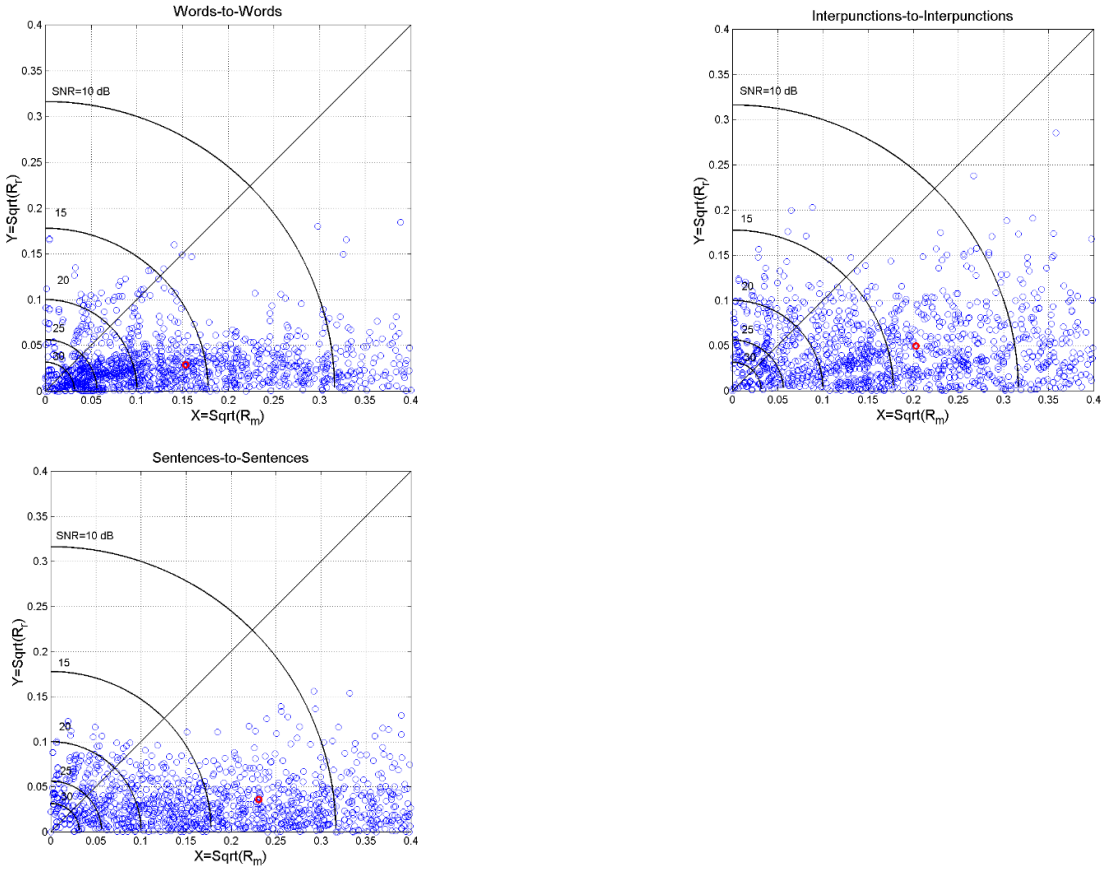


(**a**) (**b**) (**c**)

Figure 16. Mean value (upper panel) and correlation coefficient (lower panel) in the indicated languages, assuming English as reference language (to be translated) in the channels: **(a)** words–to–words; **(b)** interpunctions–to–interpunctions; **(c)** sentences–to–sentences.

This conclusion is visually evident in the scatterplot between $X = \sqrt{R_m}$ and $Y = \sqrt{R_r}$ shown in Figure 17, where a constant value of Γ traces an arch of circle [34]. It is clear that $\sqrt{R_m} \gg \sqrt{R_r}$, in other words, in the three cross channels, Γ is dominated by R_m, in agreement with what shown in Figures 15, 16.

 doi:10.20944/preprints202508.1588.v1



(**a**) (**b**) (**c**)

Figure 17. Scatterplot between $X = \sqrt{R_m}$ and $Y = \sqrt{R_r}$ in the indicated channels: **(a)** words–to–words; **(b)** interpunctions–to–interpunctions; **(c)** sentences–to–sentences. Red circles indicate the coordinates $< \sqrt{R_m} >$, $< \sqrt{R_r} >$ of the barycenter.

Finally, Figure 18 shows mean value and standard deviation of Γ in the three channels, by assuming the language indicated in abscissa as independent language/translation.

Notice that, overall, the probability distribution of Γ can be modelled as Gaussian with mean value and standard deviation reported in Table 4 (for its calculation see Appendix C). Notice that cross channels have larger Γ than the series channels (Table 3), because "translation" between two modern languages use mostly deterministic channels.

Table 4. Mean and standard deviation of the signal–to–noise ratio Γ (dB) in the indicated cross channels. The probability density function of each channel is modelled as Gaussian.

Channel	Mean Γ (dB)	Standard Deviation (dB)
Words–to–Words	18.93	9.21
Interpunctions–to–Interpunctions	15.60	7.99
Sentences–to–Sentences	14.94	8.08

 doi:10.20944/preprints202508.1588.v1

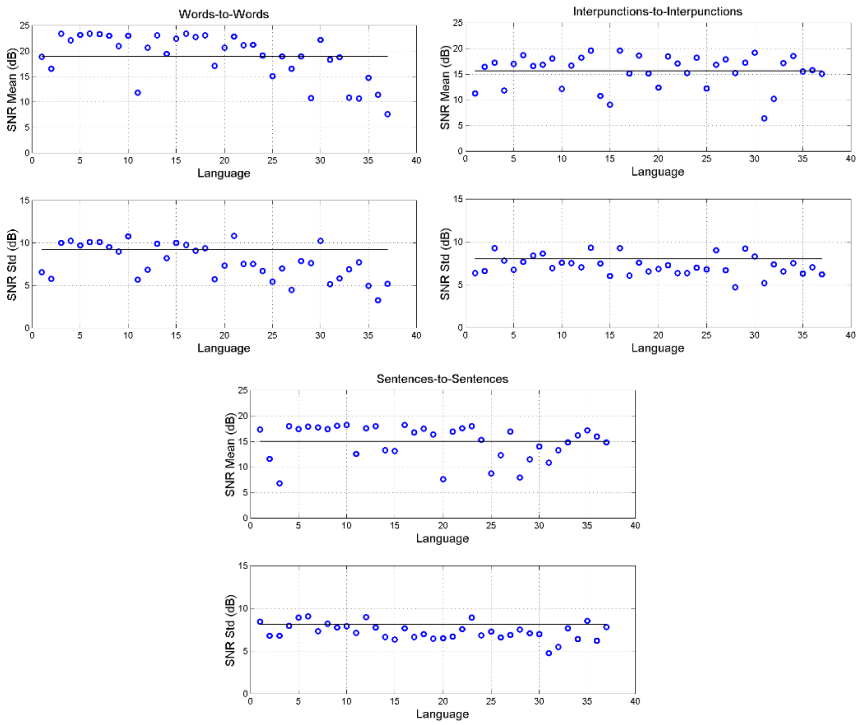


Figure 18. Mean value (upper panel) and standard deviation lower of SNR Γ (dB), in the indicated language (see Table 1), in the indicated channels: **(a)** words–to–words; **(b)** interpunctions–to–interpunctions; **(c)** sentences–to–sentences. Black lines indicate overall means. The mean of standard deviations is calculated from the mean of variances.

Figure 19 shows, as example, the modelling of the words–to–words overall channel.

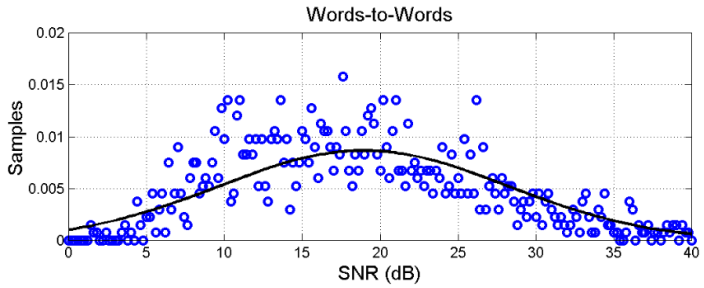


Figure 19. Histogram of the signal–to–noise ratio SNR Γ (dB) in the words–to–words channel ($37 \times 37 - 37 = 1332$ samples), blue circles. The continuous black line models the histogram with a Gaussian density function.

Now, we conjecture the characteristics of the three channels for an indistinct human being, by merging all values, as done with words in Figure 19.

Figure xx shows the Gaussian probability density functions and their probability distributions of the overall Γ in the three channels calculated with the values of Table 4. These distributions refer, therefore, to channels in which all languages merge in a single digital code. In other words, we might consider these probability distributions as "universal", typical of humans using plain text.

From Figures 18–20 and Table 4, it clearly emerges the following "universal" characteristics.

a) The words–to–words channel is distinguished from the other two channels, with larger Γ. This channel is the most deterministic.

 doi:10.20944/preprints202508.1588.v1

b) The interpunctions–to–interpunctions and sentences–to–sentences channels are very similar both in mean value and standard deviation of Γ, therefore indicating a similar freedom in creating variations with respect to their deterministic channels.

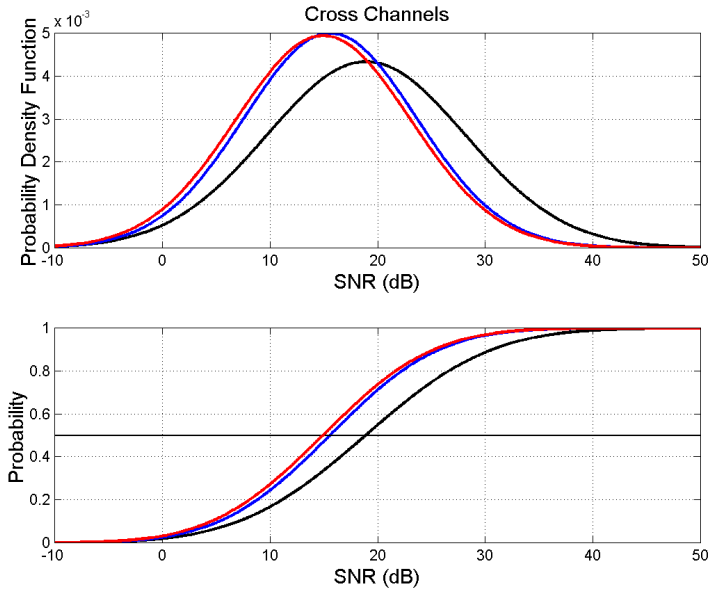


Figure 20. "Universal" Gaussian probability density function (upper panel) and probability distribution function (that the abscissa is not exceeded) of Γ (dB) in the following channels: words–to–words, black; interpunctions–to–interpunctions, blue; sentences–to–sentences, red. The horizontal black line indicates the mean value.

9. Summary and Conclusions

How the human brain analyzes the parts of a sentence (parsing) and describes their syntactic roles is still a major question in cognitive neuroscience. In References [2,3,33], we proposed that a sentence is elaborated by the short–term memory with three independent processing units in series: (1) syllables and characters to make a word, (2) words and interpunctions to make a word interval; (3) word intervals to make a sentence.

This approach is simple but useful, because the multiple processing of the brain regarding speech/text is not yet fully understood but characters, words and interpunctions – these latter needed to distinguish word intervals and sentences – can be easily studied in any alphabetical language and epoch. Our conjecture, therefore, is that we can find clues on the performance of the mind, at a high cognitive level, by studying the most abstract human invention, namely the alphabetical texts.

The aim of the present paper was to further develop and complete the theory proposed in Reference [2,3,33], and then apply it to the flow of linguistic variables making a sentence, namely, the transformation of: (a) characters into words; (b) words into word intervals; (c) word intervals into sentences. Since the connection between these linguistic variables is described by regression lines, we have analyzed experimental scatterplots between the variables.

In the first part of the article, we have recalled and further developed the theory of linear channels, which models stochastic variables linearly connected. The theory is applicable to any field/specialty in which a linear relationship holds between two variables.

We have first studied how the output variable y_k of channel k relates to the output variable y_j of another similar channel j for the same input x. These channels can be termed as "cross channels" and are fundamental in studying language translation.

Secondly, we have studied how the output of a deterministic channel relates to the output of its noisy version. A deterministic channel is not "deterministic" for what concerns the number of concepts, because, for example, the same number of sentences can communicate different meanings by just changing words and interpunctions. What is "deterministic" is the size of the ensemble.

 doi:10.20944/preprints202508.1588.v1

Then, we have studied a channel made of series of single channels and have established that its noise–to–signal ratio Γ is proportional to the average of the single channel noise–to–signal ratios.

In the second part of the article, we have explored, experimentally, the linear relationships between characters, words, interpunctions and sentences, in a large set of the New Testament books. We have considered the original Greek texts and their translation to Latin and to 35 modern languages because, in any language, they tell the same story, therefore it is meaningful to compare their translations. Moreover, they use common words, therefore, they can give some clues on how most humans communicate.

The characters–to–words channel is the nearest to be purely deterministic. It does not tend to be typical of a particular text/writer but more of a language because a writer has very little freedom in using words of very different length.

On the contrary, the channels words–to–interpunctions and the interpunctions–to–sentences are less deterministic, they depend more on writer/text than on language.

The signal–to–noise ratio Γ is as a figure of merit of the deterministic channel. The larger Γ is, the more the channel is deterministic.

In conclusion, humans have invented codes whose sequences of symbols making words cannot variate very much for indicating single physical or mental objects of their experience. On the contrary, to communicate concepts, a large variability is achieved by introducing interpunctions to make word intervals and word intervals to make sentences, the final depositary of human basic concepts. Future work should be devoted to non–alphabetical laguages.

Funding: This research received no external funding

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable

Data Availability Statement: Data are contained within the article.

Acknowledgments: The author thanks Lucia Matricciani for drawing Figures 1–5.

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

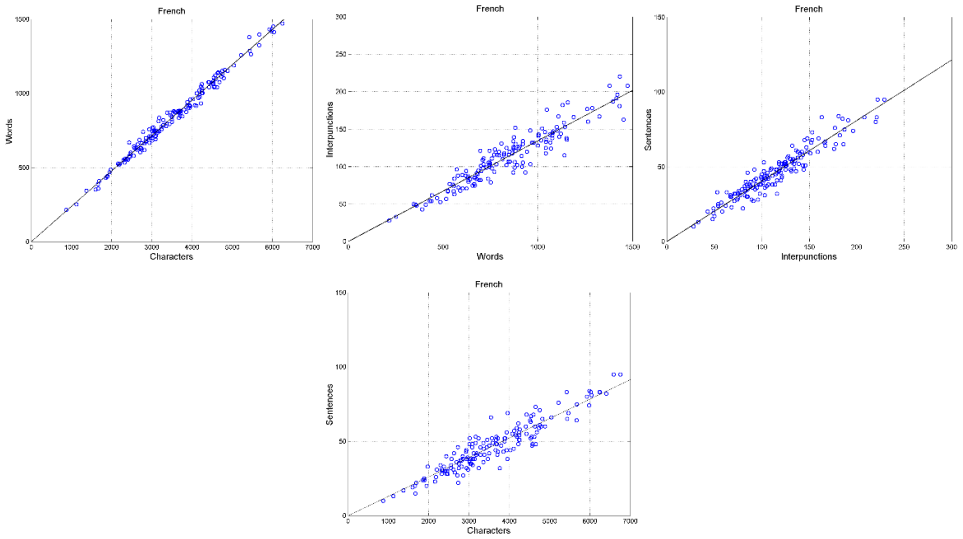
Appendix A. List of mathematical symbols

Symbol	Definition
m	Slope of regression line
m_{kj}	Slope in cross channel
m_{kj}	Correlation Coefficient in cross channel
n_C	Number of characters per chapter
n_W	Number of words per chapter
n_S	Number of sentences per chapter
n_I	Number of interpunctions per chapter
r	Correlation coefficient of linear variables
r^2	Coefficient of determination
s	Standard deviation
s^2	Variance
C_P	Characters per word
I_p	Word interval
M_F	Word intervals per sentence
N_m	Regression noise power
N_r	Correlation noise power

 doi:10.20944/preprints202508.1588.v1

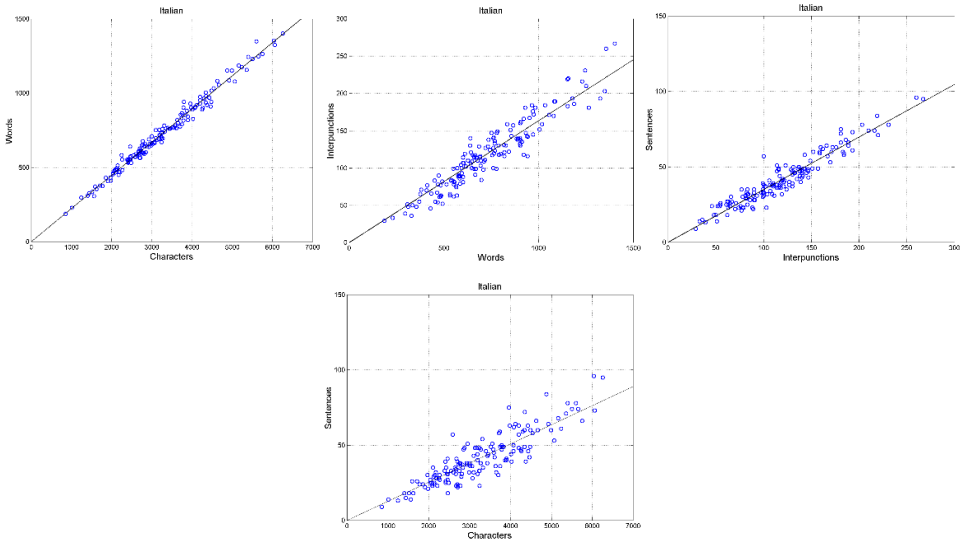
R_m	Regression noise–to–signal power ratio
R_r	Correlation noise–to–signal power ratio
P_F	Words per sentence
γ	Signal–to–noise ratio (linear)
Γ	Signal–to–noise ratio (dB)
<>	Mean value

Appendix B: Scatterplots in Different Languages



(**a**) (**b**) (c) (d)

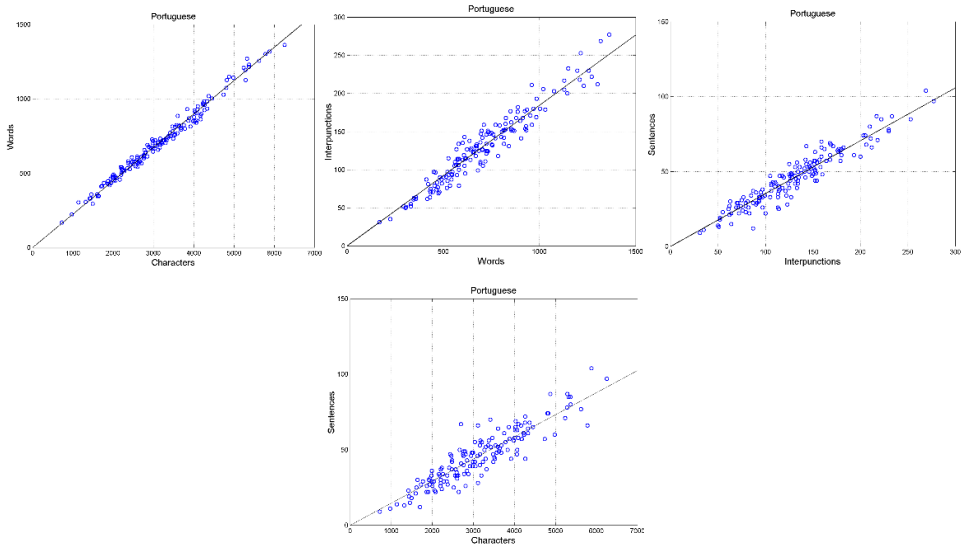
Figure A1. Scatterplots in the French translation between: (a) characters and words; (b) words and interpunctions; (c) interpunctions and sentences; (d) between characters and sentences.



(**a**) (**b**) (c) (d)

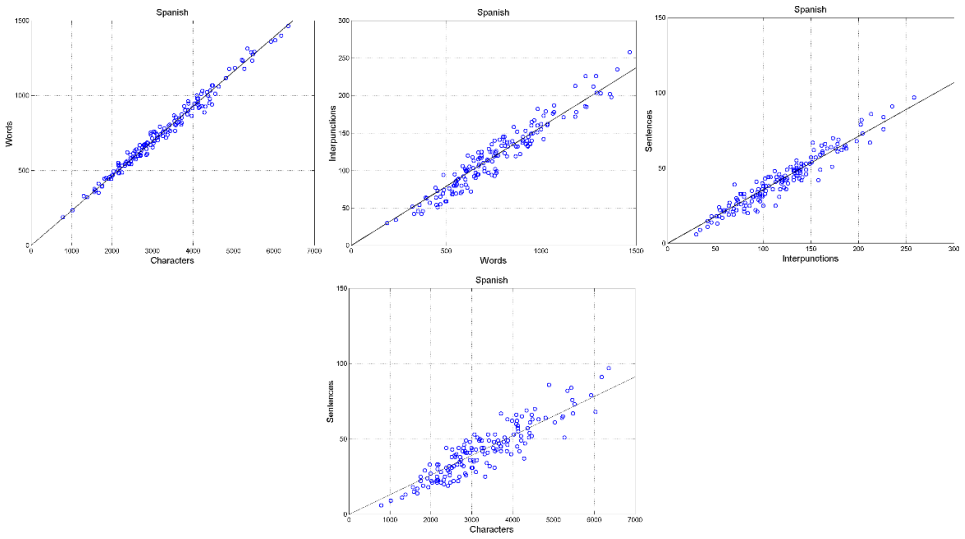
Figure A2. Scatterplots in the Italian translation between: (a) characters and words; (b) words and interpunctions; (c) interpunctions and sentences; (d) between characters and sentences.

 doi:10.20944/preprints202508.1588.v1



(**a**) (**b**) (c) (d)

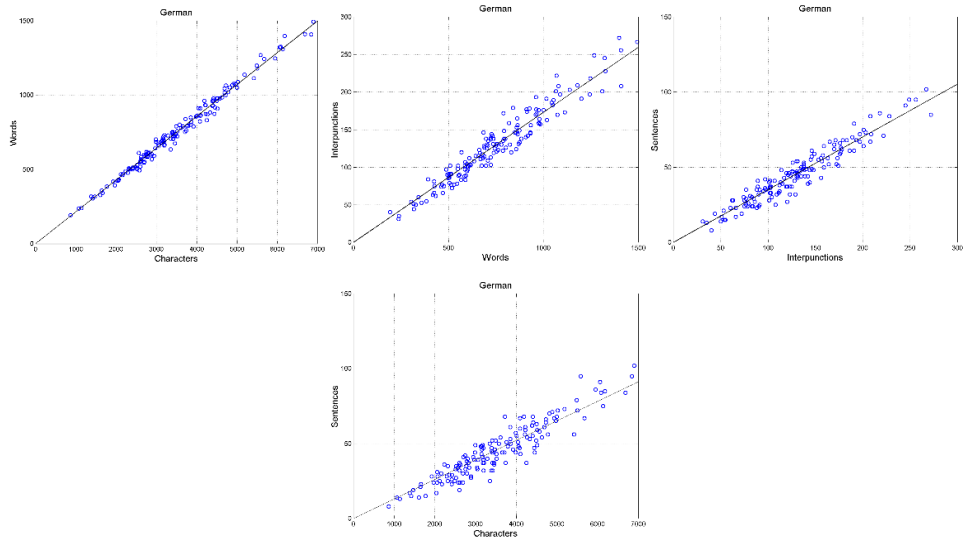
Figure A3. Scatterplots in the Portuguese translation between: (a) characters and words; (b) words and interpunctions; (c) interpunctions and sentences; (d) between characters and sentences.



(**a**) (**b**) (c) (d)

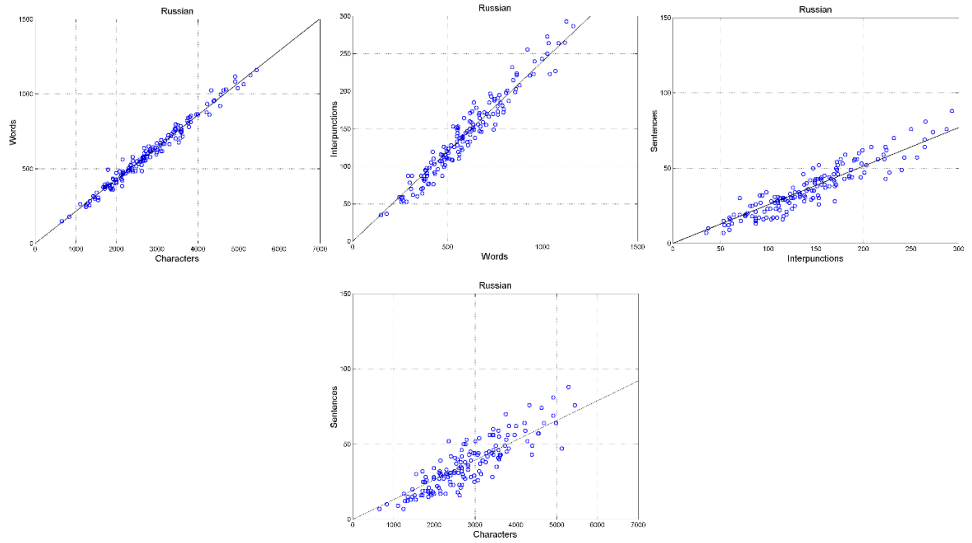
Figure A4. Scatterplots in the Spanish translation between: (a) characters and words; (b) words and interpunctions; (c) interpunctions and sentences; (d) between characters and sentences.

 doi:10.20944/preprints202508.1588.v1



(**a**) (**b**) (c) (d)

Figure A5. Scatterplots in the German translation between: (a) characters and words; (b) words and interpunctions; (c) interpunctions and sentences; (d) between characters and sentences.



(**a**) (**b**) (c) (d)

Figure A6. Scatterplots in the Russian translation between: (a) characters and words; (b) words and interpunctions; (c) interpunctions and sentences; (d) between characters and sentences.

Appendix C

Let m_k and s_k be the (conditional) mean value and standard deviation of samples belonging to set $k-th$, out of N sets of the ensemble, e.g. the values shown in Figure xx. From statistical theory [86–88], the unconditional mean (ensemble mean) m is given by the mean of means:

$$m = \frac{1}{N}\sum_{k=1}^{N} m_k \qquad \text{(A1)}$$

The unconditional variance (ensemble variance) s^2 (s is the unconditional standard deviation) is given:

$$s^2 = \mathrm{var}(m_k) + \frac{1}{N}\sum_{k=1}^{N} s_k^2 \qquad \text{(A2)}$$

$$\mathrm{var}(m_k) = \frac{1}{N}\sum_{k=1}^{N} m_k^2 - m^2 \qquad \text{(A3)}$$

From Eqs. (A1)–A(3), we get the overall values reported in Table 4.

 doi:10.20944/preprints202508.1588.v1

References

1. Deniz, F.; Nunez–Elizalde, A.O.; Huth, A.G.; Gallant Jack, L. The Representation of Semantic Information Across Human Cerebral Cortex During Listening Versus Reading Is Invariant to Stimulus Modality **2019**, *J. Neuroscience*, 39, 7722–7736.
2. Matricciani, E. A Mathematical Structure Underlying Sentences and Its Connection with Short–Term Memory. *AppliedMath* **2024**, *4*, 120–142. https://doi.org/10.3390/appliedmath4010007
3. 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.
4. Miller, G.A. The Magical Number Seven, Plus or Minus Two. Some Limits on Our Capacity for Processing Information, **1955**, *Psychological Review*, 343–352.
5. Crowder, R.G. Short–term memory: Where do we stand?, **1993**, *Memory & Cognition*, 21, 142–145.
6. Lisman, J.E., Idiart, M.A.P. Storage of 7 ± 2 Short–Term Memories in Oscillatory Subcycles, **1995,** *Science*, 267, 5203,. 1512– 1515.
7. Cowan, N., The magical number 4 in short–term memory: A reconsideration of mental storage capacity, *Behavioral and Brain Sciences*, **2000**, 87–114.
8. 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
9. Saaty, T.L., Ozdemir, M.S., Why the Magic Number Seven Plus or Minus Two, *Mathematical and Computer Modelling*, **2003**, 233–244.
10. 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.
11. Richardson, J.T.E, Measures of short–term memory: A historical review, **2007**, *Cortex*, 43, 5, 635–650.
12. Mathy, F., Feldman, J. What's magic about magic numbers? Chunking and data compression in short–term memory, *Cognition*, **2012**, 346–362.
13. Gignac, G.E. The Magical Numbers 7 and 4 Are Resistant to the Flynn Effect: No Evidence for Increases in Forward or Backward Recall across 85 Years of Data. *Intelligence*, **2015**, 48, 85–95. https://doi.org/10.1016/j.intell.2014.11.001
14. Trauzettel–Klosinski, S., K. Dietz, K. Standardized Assessment of Reading Performance: The New International Reading Speed Texts IreST, *IOVS* **2012**, 5452–5461.
15. 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.
16. Atkinson, R.C., Shiffrin, R.M., The Control of Short–Term Memory, **1971**, *Scientific American*, 225, 2, 82–91.
17. Murdock, B.B. Short–Term Memory, **1972**, *Psychology of Learning and Motivation*, 5, 67–127.
18. 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.
19. 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.
20. Grondin, S. A temporal account of the limited processing capacity, *Behavioral and Brain Sciences*, **2000**, 24, 122–123.
21. Pothos, E.M., Joula, P., Linguistic structure and short–term memory, *Behavioral and Brain Sciences*, **2000**, 138–139.
22. 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
23. 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.
24. Barrouillest, P., Camos, V., As Time Goes By: Temporal Constraints in Working Memory, *Current Directions in Psychological Science*, **2012**, 413–419.
25. Potter, M.C. Conceptual short–term memory in perception and thought, **2012**, *Frontiers in Psychology*, doi: 10.3389/fpsyg.2012.00113.

 doi:10.20944/preprints202508.1588.v1

26. Jones, G, Macken, B., Questioning short−term memory and its measurements: Why digit span measures long−term associative learning, *Cognition*, **2015**, 1−13.
27. Chekaf, M., Cowan, N., Mathy, F., Chunk formation in immediate memory and how it relates to data compression, *Cognition*, **2016**, 155, 96−107.
28. Norris, D., Short–Term Memory and Long–Term Memory Are Still Different, **2017**, *Psychological Bulletin*, 143, 9, 992–1009.
29. Houdt, G.V., Mosquera, C., Napoles, G., A review on the long short–term memory model, **2020**, *Artificial Intelligence Review*, 53, 5929–5955.
30. 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.
31. Rosenzweig, M.R., Bennett, E.L., Colombo, P.J., Lee, P.D.W. Short–term, intermediate–term and Long–term memories, , *Behavioral Brain Research*, **1993**, 57, 2, 193–198.
32. Kaminski, J. Intermediate–Term Memory as a Bridge between Working and Long–Term Memory, *The Journal of Neuroscience*, 2017, 37(20), 5045–5047.
33. Matricciani, E. Equivalent Processors Modelling the Short–Term Memory. *Preprints*, **2025**, 2025061906. https://doi.org/10.20944/preprints202506.1906.v1.
34. 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 Journal of Statistics* **2019**, *9*, 373–406. https://doi.org/10.4236/ojs.2019.93026
35. 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.
36. Matricciani, E. (2022) Multiple Communication Channels in Literary Texts. *Open Journal of Statistics*, **12**, 486–520. doi: 10.4236/ojs.2022.124030.
37. Strinati E. C., Barbarossa S. 6G Networks: Beyond Shannon Towards Semantic and Goal–Oriented Communications, **2021**, *Computer Networks*, 190, 8, 1–17.
38. Shi, G., Xiao, Y, Li, Xie, X. From semantic communication to semantic–aware networking: Model, architecture, and open problems, **2021**, *IEEE Communications Magazine*, 59, 8, 44–50.
39. Xie, H., Qin, Z., Li, G.Y., Juang, B.H. Deep learning enabled semantic communication systems, **2021**, *IEEE Trans. Signal Processing*, 69, 2663–2675.
40. Luo, X., Chen, H.H., Guo, Q. Semantic communications: Overview, open issues, and future research directions, **2022**, *IEEE Wireless Communications*, 29, 1, 210–219.
41. Wanting, Y., Hongyang, D, Liew , Z. Q., Lim , W.Y.B., Xiong, Z., Niyato, D., Chi , X., Shen , X., Miao , C. Semantic Communications for Future Internet: Fundamentals, Applications, and Challenges, **2023**, *IEEE Communications Surveys & Tutorials*, 25, 1, 213–250.
42. Xie, H., Qin, Z., Li, G.Y., Juang, B.H. Deep learning enabled semantic communication systems, **2021**, *IEEE Trans. Signal Processing*, 69, 2663–2675.
43. Bellegarda, J.R., Exploiting Latent Semantic Information in Statistical Language Modeling, **2000**, *Proceedings of the IEEE*, 88, 8, 1279–1296.
44. D'Alfonso, S., On Quantifying Semantic Information, **2011**, *Information*, 2, 61–101; doi:10.3390/info2010061
45. Zhong, Y., A Theory of Semantic Information, **2017**, *China Communications*, 1–17.
46. Papoulis Papoulis, A. *Probability & Statistics*; Prentice Hall: Hoboken, NJ, USA, **1990**.
47. Matricciani, E. Domestication of Source Text in Literary Translation Prevails over Foreignization. *Analytics* **2025**, *4*, 17. https://doi.org/10.3390/analytics4030017