Scientific Research Publishing

Open Journal of Statistics, 2020, 10, 936-997
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

A Statistical Theory of Language Translation Based on Communication Theory

Emilio Matricciani

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy
Email: Emilio.Matricciani@polimi.it

How to cite this paper: Matricciani, E. (2020) A Statistical Theory of Language Translation Based on Communication Theory. *Open Journal of Statistics*, **10**, 936-997.
https://doi.org/10.4236/ojs.2020.106055

Received: October 29, 2020
Accepted: December 4, 2020
Published: December 7, 2020

Abstract

We propose the first statistical theory of language translation based on communication theory. The theory is based on New Testament translations from Greek to Latin and to other 35 modern languages. In a text translated into another language, all linguistic variables do numerically change. To study the chaotic data that emerge, we model any translation as a complex communication channel affected by "noise", studied according to Communication Theory applied for the first time to this channel. This theory deals with aspects of languages more complex than those currently considered in machine translations. The input language is the "signal", the output language is a "replica" of the input language, but largely perturbed by noise, indispensable, however, for conveying the meaning of the input language to its readers. We have defined a noise-to-signal power ratio and found that channels are differently affected by translation noise. Communication channels are also characterized by channel capacity. The translation of novels has more constraints than the New Testament translations. We propose a global readability formula for alphabetical languages, not available for most of them, and conclude with a general theory of language translation which shows that direct and reverse channels are not symmetric. The general theory can also be applied to channels of texts belonging to the same language both to study how texts of the same author may have changed over time, or to compare texts of different authors. In conclusion, a common underlying mathematical structure governing human textual/verbal communication channels seems to emerge. Language does not play the only role in translation; this role is shared with reader's reading ability and short-term memory capacity. Different versions of New Testament within the same language can even seem, mathematically, to belong to different languages. These conclusions are everlasting because valid also for ancient Roman and Greek readers.

Keywords

Channel Capacity, Communication Theory, Greek, Latin, Linguistic Variables, Modern Languages, New Testament, Noise-to-Signal Power Ratio, Readability Index, Short-Term Memory Capacity, Symmetry

1. A Communication Channel Approach to the Theory of Translation

Translation is the replacement of textual material in one language by equivalent textual material in another language. It transfers meaning from one set of patterned symbols into another set of patterned symbols. Translation was formerly studied as a language-learning methodology or as part of comparative literature. Over time, however, the interdisciplinary and specialization of the subject have become more evident and theories and models have continued to be imported from other disciplines [1] [2]. References [3]-[9] report results not based on mathematical analysis of texts, as we do with the theory here proposed. When a mathematical approach is used, as in References [10]-[25], most of these studies neither concern the aspects of Shannon's communication theory [26], nor the fundamental connection which some linguistic variables have with reader's reading ability and short-term memory capacity, considered instead in this paper. In fact, these studies are mainly concerned with machine translations, not with a response of human readers. Very often they refer only to one linguistic variable, e.g. phrases [24]. As stated in [25], statistical machine translation is a process in which the text to be translated is "decoded" by eliminating the noise by adjusting lexical and syntactic divergences to reveal the intended message. In this paper, on the contrary, what we define as "noise"—given by quantitative differences between source text and translated text—must not be eliminated because it makes the translation readable and matched to reader's short-term memory capacity, a connection never considered in the mentioned references.

The aim of this paper is to show that there seems to be a mathematical/statistical background that unifies all alphabetical languages, despite the spreading of the statistics of linguistic variables from language to language, described by parallel communication channels, one for each linguistic variable. The differences between translations seem to be mostly due to differences in the expected reader's reading ability—quantified by readability formulae—and reader's short-term memory capacity—quantified by Miller's 7 ± 2 law [27]—assumed by the translators, not to a particular language. In other words, it seems that the mythical biblical Tower of Babel has produced a lot of "noise", but has not destroyed this common background.

This unifying picture is mainly assessed by defining firstly an ideal translation channel, and secondly by comparing the actual translation channels to it, according to communication theory. Shannon has set the fundamental mathemat-

ics of the main parts of a communication channel [26]: the source of information (input) and the channel to which this information is delivered, with its response (output) to the input. The source is characterized by its entropy, *i.e.* the minimum average number of bits necessary for coding a symbol randomly produced by the source; the channel is characterized by the signal-to-noise ratio, which determines its capacity (in bits per symbol).

In this paper, we study the translation channel, after suitably defining input and output symbols. Compared to Shannon's channels, our linguistic channels work at a different level because, for equal meaning, both input and output texts are structured in such a way to match reader's expected reading ability and short-term memory capacity. In other words, these channels do not communicate with a machine, but with human beings, who may have serious difficulties in understanding what they read, if the text is not matched to their own reading ability and short-term memory capacity. The peculiarity of these linguistic channels makes less important the translation language. None of the previous studies has considered this unified approach.

The main mathematical/statistical characteristics are determined by studying the translation of a large selection of New Testament (NT) books—namely *Matthew*, *Mark*, *Luke*, *John*, *Acts*, *Epistle to the Romans*, *Apocalypse*, for a total of 155 chapters, according to the traditional subdivision of the original Greek texts—from Greek to Latin and to other 35 modern languages, 36 translations in total. The theory does not include meaning.

The rationale for studying the NT translations is based on their accuracy in any translation because done by a team of experts, whose aim is to render the same meaning of the original Greek texts to their readers, regardless of the language used. These translations strictly respect the subdivision in chapters and verses of the Greek texts, therefore they can be studied at least at these two different levels (chapters and verses), by comparing how a variable varies quantitatively from translation to translation. Notice that "translation" should not be confused with "language" because language plays one of the roles in translations, not the only one. For our analysis, we have chosen the chapter level because the amount of text is sufficiently large to assess reliable statistics. Therefore, for each translation we have considered a database of $155 \times 37 = 5735$ samples for each stochastic variable, sufficiently large to give reliable results.

After this introduction, the rest of the paper is organized as follows. In Section 2 we list the NT translations and some fundamental statistics; in Section 3 we define the ideal translation and the real translation; in Section 4 we define the communication channel, its noise-to-signal power ratio and describe its geometrical representation; in Section 5 we define the linguistic communication channels and study them; in Section 6 we deal with channel capacity according to communication theory; in Section 7 we relate the number of words per interpunctions (also termed the word interval) to human short-term memory capacity; in Section 8 we define a readability formula applicable to any alphabetical

language; in Section 9 we examine different versions of the NT translation in the same language; in Section 10 we compare the translations of a novel from English to some modern languages and compare their statistics with the NT translation statistics from English to the same languages; in Section 11 we propose a general theory of language translation; finally, in Section 12 we summarize the main results and draw some conclusions. Some appendices report more details.

2. Translations: Babel of Different Statistical Results

Following our statistical study of a large corpus of literary texts taken from the Italian Literature spanning seven centuries [28], we use the same stochastic variables to study the NT translations, namely, the number of words n_W, sentences n_S and interpunctions n_I per chapter, the number of characters per word C_P, the number of words per sentence P_F, the number of words per interpunctions I_P—a variable that seems to be related to the short-term memory capacity [28]—the number of interpunctions per sentence M_F, which gives also the number of word intervals per sentence, and the total number of words W, sentences S and interpunctions I (Appendix A lists all mathematical symbols). How interpunctions were inserted into the *scriptio continua* of Greek and Latin texts is reviewed and discussed in [29].

Table 1 lists the NT translations considered in our study, with some first-order statistics. We have considered only alphabetical languages, listed according to their linguistic family for visualizing possible similarities. Esperanto is an artificial (constructed) language based on European languages.

We downloaded each translation text from the web sites reported in Table 1, and saved the text in WinWord format. Then, for each chapter we counted words, sentences and interpunctions (full-stops, question marks, exclamation marks, commas, colons, semicolons) after deleting all extraneous characters added to the original text by translators/commentators, such as titles, footnotes et cetera. At the end of this lengthy and laborious work, only the original text *sine glossa* was left to be studied. Of course, we do not need to understand any of the translation languages because the process consists in just counting characters and sequences of characters.

The first impression arising after reading these statistics is their large variety. Words, sentences and their distribution within chapters (for sentences) and within sentences (for words) can be very different from translation to translation. Even though all these texts convey the same meaning, the spread—*i.e.* the scattering of the values—is large. For example, the number of total words ranges from 90,799 (Latin) to 152,823 (Haitian), a spread of 62,024 words which represents 61.9% of the total number of words in Greek, 100,145; the number of total sentences ranges from 5370 (Latin) to 10,429 (Haitian), a spread of 106.3% of the total number of sentences in Greek, 4759. However, a ranking is evident as some translations are closer to the Greek originals than others. A similar spread is also noticeable in the average and standard deviation of the number of words

per sentence P_F (from Hebrew to Welsh, range 52.4%), words per interpunction I_P—word interval—(from Russian to Cebuano, range 60.8%) and interpunctions per sentence M_F (from Cebuano to Esperanto, range 79.5%), Table 2.

Table 1. List of languages used in the NT translations (*Matthew*, *Mark*, *Luke*, *John*, *Acts*, *Epistle to the Romans*, *Apocalypse*). Total number of words W, sentences S and interpunctions I. Average values of characters per word C_P, and words n_W, sentences n_S, interpunctions n_I per chapter. In brackets: standard deviation. Last access to the indicated web sites was in the week October 5 to 9, 2020.

Language	Family	W	C_P	n_W	S	n_S	I	n_I
Greek[1]	Hellenic	100,145	4.86 (0.25)	646.1 (220.4)	4759	30.7 (14.0)	13,698	88.4
Latin[2]	Italic	90,799	5.16 (0.28)	585.8 (206.5)	5370	34.6 (15.9)	18,380	118.6
Esperanto[3]	Constructed	111,259	4.43 (0.20)	717.8 (245.6)	5483	35.4 (15.6)	22,552	145.5
French[4]	Romance	133,050	4.20 (0.16)	858.4 (282.5)	7258	46.8 (17.0)	18,284	118.0
Italian[5]	Romance	112,943	4.48 (0.19)	728.7 (246.3)	6396	41.7 (16.7)	17,904	115.5
Portuguese[6]	Romance	117,537	4.43 (0.20)	706.2 (239.6)	6518	45.7 (18.1)	18,410	118.8
Romanian[7]	Romance	109,468	4.34 (0.19)	766.1 (265.4)	7080	45.3 (19.5)	20,105	129.7
Spanish[8]	Romance	118,744	4.30 (0.19)	758.3 (252.3)	7021	42.1 (17.4)	18,587	119.9
Danish[9]	Germanic	131,021	4.14 (0.16)	845.3 (299.1)	8762	56.5 (22.5)	22,196	143.2
English[10]	Germanic	122,641	4.24 (0.17)	791.2 (274.8)	6590	42.5 (17.3)	16,666	107.5
Finnish[11]	Germanic	95,879	5.90 (0.31)	618.6 (216.6)	5893	38.0 (16.9)	19,725	127.3
German[12]	Germanic	117,269	4.68 (0.19)	756.6 (258.0)	7069	45.6 (18.4)	20,233	130.5
Icelandic[13]	Germanic	109,170	4.34 (0.18)	704.3 (243.2)	7193	46.4 (18.5)	19,577	126.3
Norwegian[14]	Germanic	140,844	4.08 (0.13)	908.7 (313.3)	9302	60.0 (20.8)	18,370	118.5
Swedish[15]	Germanic	118,833	4.23 (0.18)	766.7 (268.9)	7668	49.5 (19.6)	15,139	97.7
Bulgarian[16]	Balto-Slavic	111,444	4.41 (0.19)	719.0 (246.8)	7727	49.8 (20.1)	20,093	129.6
Czech[17]	Balto-Slavic	92,533	4.51 (0.21)	597.0 (203.0)	7514	48.5 (21.2)	19,465	125.6
Croatian[18]	Balto-Slavic	97,336	4.39 (0.22)	628.0 (220.6)	6750	43.6 (18.7)	17,698	114.2
Polish[19]	Balto-Slavic	99,592	5.10 (0.22)	642.5 (224.6)	8181	52.8 (18.9)	21,560	139.1
Russian[20]	Balto-Slavic	92,736	4.67 (0.27)	598.3 (208.3)	5532	36.1 (16.4)	22,083	142.5
Serbian[21]	Balto-Slavic	104,585	4.24 (0.20)	674.7 (237.0)	7532	48.6 (20.0)	18,251	117.7
Slovak[22]	Balto-Slavic	100,151	4.65 (0.23)	646.1 (223.7)	8023	51.8 (20.8)	19,690	127.0
Ukrainian[23]	Balto-Slavic	107,047	4.56 (0.26)	690.6 (247.9)	8043	51.9 (21.0)	22,761	146.8
Estonian[24]	Uralic	101,657	4.89 (0.24)	655.9 (229.8)	6310	40.7 (17.5)	19,029	122.8
Hungarian[25]	Uralic	95,837	5.31 (0.29)	618.3 (212.3)	5971	38.5 (16.7)	22,970	148.2
Albanian[26]	Albanian	123,625	4.07 (0.22)	797.6 (278.1)	5807	37.5 (16.4)	19,352	124.9
Armenian[27]	Armenian	100,604	4.75 (0.40)	649.1 (235.6)	6595	42.6 (18.7)	18,086	116.7

Continued

Welsh[28]	Celtic	130,698	4.04 (0.15)	843.2 (299.6)	5676	36.6 (15.8)	22,585	262.4
Basque[29]	Isolate	94,898	6.22 (0.27)	612.2 (219.7)	5591	36.1 (15.8)	19,312	124.6
Hebrew[30]	Semitic	88,478	4.22 (0.17)	570.8 (199.7)	7597	49.0 (20.4)	15,806	102.0
Cebuano[31]	Austronesian	146,481	4.65 (0.10)	945.0 (326.6)	9221	59.5 (22.4)	16,788	108.3
Tagalog[32]	Austronesian	128,209	4.83 (0.17)	827.2 (283.6)	7944	51.2 (21.2)	16,405	105.8
Chichewa[33]	Niger-Congo	94,817	6.08 (0.18)	611.7 (203.7)	7560	48.8 (18.3)	15,817	102.0
Luganda[34]	Niger-Congo	91,819	6.23 (0.23)	592.4 (207.9)	7073	45.6 (18.8)	16,401	105.8
Somali[35]	Afro-Asiatic	109,686	5.32 (0.16)	707.7 (236.1)	6127	39.5 (17.9)	17,765	114.6
Haitian[36]	French Creole	152,823	3.37 (0.10)	986.0 (330.1)	10429	67.3 (24.3)	23,813	153.6
Nahuatl[37]	Uto-Aztecan	121,600	6.71 (0.24)	784.5 (260.3)	9263	59.8 (21.4)	19,271	124.3

[1]https://www.biblegateway.com/versions/Tyndale–House–Greek–New–Testament/#booklist
[2]http://www.vatican.va/archive/bible/nova_vulgata/documents/nova-vulgata_novum-testamentum_lt.html
[3]https://newchristianbiblestudy.org/bible/esperanto/
[4]https://www.bibliacatolica.com.br/
[5]http://www.vatican.va/archive/ITA0001/_INDEX.HTM
[6]https://www.bibliacatolica.com.br/
[7]https://www.biblegateway.com/versions/Nou%C4%83-Traducere-%C3%8En-Limba-Rom%C3%A2n%C4%83-NTLR/#booklist
[8]http://www.vatican.va/archive/ESL0506/_INDEX.HTM
[9]https://www.biblegateway.com/versions/Bibelen-p%C3%A5-hverdagsdansk-BPH/#booklist
[10]http://www.vatican.va/archive/ENG0839/_INDEX.HTM
[11]https://www.biblegateway.com/versions/Raamattu-1933-1938/#booklist
[12]https://www.biblegateway.com/versions/Raamattu-1933-1938/#booklist
[13]https://www.uibk.ac.at/theol/leseraum/bibel/mt1.html
[14]https://www.biblegateway.com/versions/Icelandic-Bible/#booklist
[15]https://www.biblegateway.com/versions/En-Levende-Bok-LB/#booklist
[16]https://www.biblegateway.com/versions/nuBibeln-Swedish-Contemporary-Bible-NUB/#booklist
[17]https://www.biblegateway.com/versions/Bulgarian-Bible-Easy-to-Read-Version-ERV-BG/#booklist
[18]https://www.biblegateway.com/versions/Bible-21-B21/#booklist
[19]https://www.biblegateway.com/versions/Hrvatski-Novi-Zavjet-Rijeka-2001-HNZ-RI/#booklist
[20]https://www.biblegateway.com/versions/Nowe-Przymierze/#booklist
[21]https://www.biblegateway.com/versions/Russian-Synodal-Version-RUSV/#booklist
[22]https://www.biblegateway.com/versions/New-Serbian-Translation-NSP-Bible/#booklist
[23]https://www.biblegateway.com/versions/N%C3%A1dej-pre-kazd%C3%A9ho-NPK/#booklist
[24]https://www.biblegateway.com/versions/Ukrainian-Bible-Easy-to-Read-Version-ERV-UK/#booklist
[25]https://newchristianbiblestudy.org/bible/estonian/
[26]https://www.biblegateway.com/versions/Hungarian-New-Translation/#booklist
[27]https://www.biblegateway.com/versions/Albanian-Bible-ALB/#booklist
[28]https://studybible.info/Armenian
[29]https://www.biblegateway.com/versions/Beibl-William-Morgan-BWM-Bible/#booklist
[30]http://www.vc.ehu.es/gordailua/testamentu.htm
[31]https://www.biblegateway.com/versions/Habrit-Hakhadasha-Haderekh/#booklist
[32]https://www.biblegateway.com/versions/Ang-Pulong-Sa-Dios-APSD-Cebuano/#booklist
[33]https://www.biblegateway.com/versions/Filipino-Standard-Version-Biblia-FSV/#booklist
[34]https://www.biblegateway.com/versions/Mawu-a-Mulungu-mu-Chichewa-Chalero-Word-of-God-in-Contemporary-Chichewa-CCL/#booklist
[35]https://www.biblegateway.com/versions/Endagaano-Enkadde-nEndagaano-Empya-Luganda-Contemporary-Bible-LCB/#booklist
[36]https://www.biblegateway.com/versions/Somali-Bible-SOM/#booklist
[37]https://www.biblegateway.com/versions/Haitian-Creole-Version-HCV/#booklist

Table 2. Words per sentence P_F, words per interpunction I_P (word interval) and interpunctions per sentence M_F (word intervals per sentence). The first number gives the average value, the number in brackets gives the standard deviation, and the third number gives the correlation coefficient between the two stochastic variables that define the parameter.

Language	P_F	I_P	M_F
Greek	23.07 (6.65) 0.897	7.47 (1.09) 0.930	3.08 (0.73) 0.938
Latin	18.28 (4.77) 0.901	5.07 (0.68) 0.952	3.60 (0.77) 0.937
Esperanto	21.83 (5.22) 0.916	5.05 (0.57) 0.967	4.30 (0.76) 0.955
French	18.73 (2.51) 0.942	7.54 (0.85) 0.948	2.50 (0.32) 0.951
Italian	18.33 (3.27) 0.907	6.38 (0.95) 0.948	2.89 (0.40) 0.954
Portuguese	16.18 (3.25) 0.913	5.54 (0.59) 0.962	2.93 (0.56) 0.948
Romanian	18.00 (4.19) 0.910	6.49 (0.74) 0.959	2.78 (0.65) 0.938
Spanish	19.07 (3.79) 0.926	6.55 (0.82) 0.962	2.91 (0.47) 0.958
Danish	15.38 (2.15) 0.935	5.97 (0.64) 0.957	2.59 (0.33) 0.955
English	19.32 (3.20) 0.917	7.51 (0.93) 0.951	2.58 (0.39) 0.948
Finnish	17.44 (4.09) 0.939	4.94 (0.56) 0.962	3.54 (0.75) 0.946
German	17.23 (2.77) 0.949	5.89 (0.60) 0.962	2.94 (0.45) 0.955
Icelandic	15.72 (2.58) 0.934	5.69 (0.67) 0.960	2.77 (0.39) 0.953
Norwegian	15.21 (1.43) 0.968	7.75 (0.84) 0.958	1.98 (0.22) 0.962
Swedish	15.95 (2.17) 0.959	8.06 (1.35) 0.922	2.01 (0.31) 0.950
Bulgarian	14.97 (2.61) 0.930	5.64 (0.64) 0.959	2.67 (0.43) 0.948
Czech	13.20 (3.10) 0.920	4.89 (0.65) 0.950	2.71 (0.61) 0.928
Croatian	15.32 (3.54) 0.928	5.62 (0.75) 0.950	2.72 (0.49) 0.961
Polish	12.34 (1.93) 0.913	4.65 (0.43) 0.965	2.67 (0.40) 0.925
Russian	17.90 (4.46) 0.898	4.28 (0.46) 0.971	4.18 (0.92) 0.927
Serbian	14.46 (2.42) 0.929	5.81 (0.69) 0.951	2.50 (0.39) 0.944
Slovak	12.95 (2.10) 0.929	5.18 (0.61) 0.953	2.51 (0.36) 0.954
Ukrainian	13.81 (2.18) 0.963	4.72 (0.41) 0.973	2.95 (0.58) 0.945
Estonian	17.09 (3.89) 0.927	5.45 (0.66) 0.956	3.14 (0.64) 0.947
Hungarian	17.37 (4.54) 0.943	4.25 (0.45) 0.972	4.09 (0.93) 0.948
Albanian	22.72 (4.86) 0.925	6.52 (0.78) 0.961	3.48 (0.61) 0.958
Armenian	16.09 (3.07) 0.930	5.63 (0.52) 0.970	2.86 (0.47) 0.964
Welsh	24.27 (4.75) 0.941	5.84 (0.44) 0.982	4.16 (0.76) 0.949
Basque	18.09 (4.31) 0.934	4.99 (0.52) 0.967	3.63 (0.81) 0.951
Hebrew	12.17 (2.04) 0.935	5.65 (0.59) 0.962	2.16 (0.33) 0.964
Cebuano	16.15 (1.71) 0.968	8.82 (1.01) 0.947	1.85 (0.22) 0.958
Tagalog	16.98 (3.24) 0.943	7.92 (0.82) 0.956	2.16 (0.44) 0.936
Chichewa	12.89 (1.79) 0.940	6.18 (0.87) 0.942	2.10 (0.25) 0.960
Luganda	13.65 (2.78) 0.931	5.74 (0.82) 0.949	2.39 (0.40) 0.950
Somali	19.57 (5.50) 0.882	6.37 (1.01) 0.930	3.06 (0.65) 0.940
Haitian	14.87 (1.83) 0.943	6.55 (0.71) 0.967	2.28 (0.26) 0.957
Nahuatl	13.36 (1.70) 0.938	6.47 (0.91) 0.930	2.08 (0.24) 0.958

For avoiding misuse of the results reported in Table 1, Table 2, notice that the average values shown in Table 2 do not coincide with averages calculable from Table 1, because, in general, the average value of a ratio is not equal to the ratio calculated from the total values. For example, for Greek the total number of words divided by the total number of sentences (*i.e.*, an estimate of the average value of the variable "words per sentence"), from Table 1 is 100,145/4759 = 21.04, while the average value of the ratio of the number of words per chapter divided by the number of sentences per chapter is 23.07 (Table 2).

The correlation r between the number of characters and the number of words is not reported because, as for Italian [28], for all languages $r > 0.990$. Finally, notice that the lists of names (Genealogy) in Matthew 1.1 - 1.17 and in Luke 3.23 - 3.38 have been deleted for not biasing the statistics of all linguistic variables. In the following sections we investigate in-depth all these variables.

3. The Ideal Translation and the Real Translation

When a text written in a language is translated into a text written in another language, all linguistic variables do numerically change. Besides the total number of words W, sentences S and interpunctions I, the other main linguistic variables are the number of words n_W, sentences n_S, and interpunctions n_I, per chapter. To them we add the number of characters per word C_P, words per sentence P_F, words per interpunctions I_P, interpunctions per sentence M_F. We refer to this latter set of variables as the deep-language variables. All these variables of language Y can be statistically compared to those of a reference language X (Greek) by calculating the correlation coefficient[38] r between any couple of variables y of language/translation Y and x of the reference language/translation X (in the following, where no confusion is possible, we refer to a variable and to the language/translation with the same mathematical symbol), and their expected regression line (*i.e.*, the relationship between averages):

$$y = mx \tag{1}$$

with m the slope of the line. Of course, we expect, and it is so in the following, that no translation can yield $r = 1$ and $m = 1$, a case referred to as the *ideal translation*. In practice, we always find $r < 1$ and $|m| \neq 1$. The slope m measures the multiplicative "bias" of the dependent variable compared to the independent variable, the correlation coefficient measures how "precise" the linear fit is. Even though the ideal translation is never found, it is useful as a reference model to which real translations can be compared. In the following we refer to it as the self-translation channel.

Figure 1 shows the scatterplot between n_W in Greek and n_W in the other languages listed in Table 1, with the regression lines (1); it shows with greater detail what reported in Table 1, Table 2. We can notice that translations can use more

[38]The correlation coefficient r between two variables x, y, with averages m_x, m_y and standard deviations s_x, s_y, is given by $r = \mu_{xy} / (s_x s_y)$, where $\mu_{xy} = \langle \{ (x - m_x)(y - m_y) \} \rangle$ is the covariance; $\langle \{ \} \rangle$ indicates average value [30].

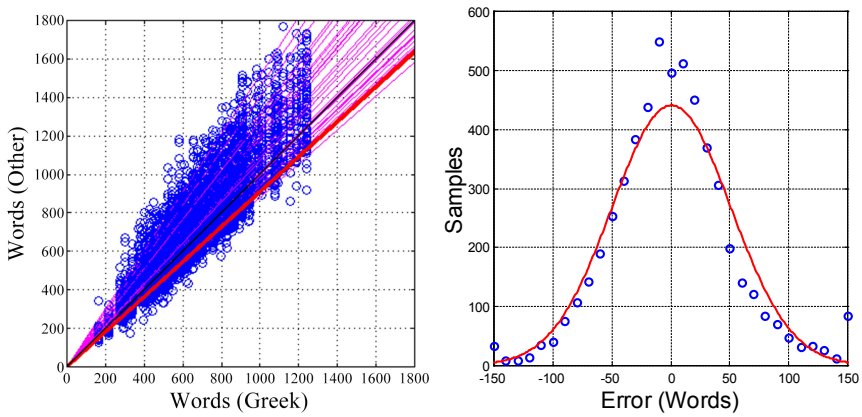


Figure 1. Left: Scatterplot between n_W in Greek and n_W in the other translations listed in Table 1, together with the regression lines (1). The black line is the line $y = x$. The red line is the regression line between Latin and Greek. **Right:** Histogram of the difference ("error") between the actual number of words in a given translation and the number of words in that translation calculated from the regression line, for a given Greek value.

or fewer words than Greek, and that Latin (red line) is one of the closest translations to Greek. Table 3 lists the values of the correlation coefficient r and slope m. Latin is the translation better correlated to Greek ($r = 0.994$), Hebrew the worst (0.949).

According to the regression lines, *i.e.*, to the relationship between the average values in Y for assigned values in X (Greek), the translations that reduce the number of words (regression lines below the 45˚ line $y = x$ in Figure 1) mostly belong to the Balto-Slavic family, while the translations that increase this number belong to the Romance and Anglo-Saxon families (except Finnish). The range is $0.881 \le m \le 1.518$. Figure 1 also shows the histogram of the difference ("error") between the actual number of words in a given translation and the average number of words in that translation calculated from the regression line, for a given Greek value. The spread of these latter values makes $r < 1$. The probability density function deducible from Figure 1 can be modelled as Gaussian.

Figure 2 shows the results concerning n_S. All languages/translations have more sentences than Greek, ranging from Latin ($m = 1.123$) to Haitian ($m = 2.085$), Table 3, therefore implying a multiplicative bias larger than the words bias, and saying that translations have very different distributions of full stops and, in general, interpunctions, not only compared to Greek, but also compared to each other. The correlation coefficients are all significantly lower than those concerning n_W, in the range $0.899 \le r \le 0.969$, Table 3. All translations convey the same meaning but with different quantities of words and sentences.

Figure 3 shows the results concerning n_I. Most translations use more interpunctions than Greek, ranging from Swedish ($m = 1.107$) to Haitian ($m = 1.730$), see Table 3, therefore implying, again, a multiplicative bias larger than that found with words and sentences. Interpunctions impact directly on readers' reading ability and short-term memory capacity. The correlation coefficient varies in the range $0.938 \le r \le 0.974$.

Table 3. Slope m and correlation coefficient r of the regression line $y = mx$ between a given stochastic variable in a translation and the corresponding variable in the original Greek text.

Language	n_W		n_S		n_I		P_F		I_P		M_F	
	m	r	m	r	m	r	m	r	m	r	m	r
Greek	1	1	1	1	1	1	1	1	1	1	1	1
Latin	0.909	0.994	1.123	0.969	1.347	0.965	0.780	0.883	0.673	0.688	1.149	0.731
Esperanto	1.110	0.991	1.139	0.966	1.647	0.961	0.925	0.842	0.668	0.586	1.356	0.621
French	1.320	0.970	1.458	0.939	1.293	0.960	0.768	0.619	0.999	0.676	0.779	0.449
Italian	1.125	0.985	1.303	0.935	1.338	0.957	0.757	0.602	0.846	0.536	0.894	0.218
Portuguese	1.091	0.984	1.442	0.948	1.459	0.954	0.675	0.728	0.732	0.502	0.922	0.546
Romanian	1.186	0.983	1.450	0.956	1.348	0.955	0.759	0.798	0.858	0.536	0.888	0.675
Spanish	1.168	0.980	1.338	0.957	1.338	0.947	0.797	0.760	0.866	0.503	0.914	0.523
Danish	1.308	0.963	1.784	0.943	1.612	0.958	0.630	0.571	0.789	0.625	0.805	0.394
English	1.225	0.986	1.346	0.942	1.216	0.966	0.797	0.647	0.995	0.597	0.808	0.456
Finnish	0.959	0.987	1.226	0.969	1.439	0.965	0.738	0.850	0.655	0.656	1.128	0.743
German	1.167	0.968	1.440	0.936	1.468	0.953	0.713	0.721	0.779	0.674	0.920	0.468
Icelandic	1.088	0.974	1.461	0.926	1.427	0.957	0.650	0.701	0.754	0.619	0.866	0.458
Norwegian	1.401	0.956	1.848	0.905	1.331	0.958	0.612	0.152	1.024	0.548	0.609	0.073
Swedish	1.187	0.980	1.559	0.939	1.107	0.960	0.656	0.700	1.073	0.673	0.631	0.565
Bulgarian	1.110	0.973	1.572	0.929	1.458	0.953	0.614	0.492	0.746	0.556	0.830	0.305
Czech	0.922	0.985	1.552	0.944	1.422	0.950	0.554	0.715	0.649	0.650	0.856	0.501
Croatian	0.974	0.992	1.393	0.956	1.291	0.967	0.646	0.802	0.747	0.720	0.864	0.711
Polish	0.995	0.986	1.631	0.899	1.558	0.959	0.498	0.179	0.614	0.600	0.824	0.124
Russian	0.927	0.990	1.164	0.952	1.610	0.958	0.756	0.769	0.566	0.542	1.327	0.646
Serbian	1.045	0.983	1.539	0.936	1.326	0.961	0.600	0.734	0.770	0.695	0.786	0.603
Slovak	0.997	0.964	1.634	0.939	1.432	0.954	0.536	0.703	0.686	0.625	0.785	0.508
Ukrainian	1.071	0.972	1.637	0.924	1.650	0.959	0.568	0.594	0.621	0.431	0.923	0.522
Estonian	1.016	0.985	1.304	0.965	1.389	0.967	0.720	0.794	0.723	0.717	0.997	0.640
Hungarian	0.956	0.986	1.235	0.956	1.671	0.961	0.734	0.740	0.561	0.512	1.301	0.650
Albanian	1.235	0.984	1.203	0.965	1.409	0.956	0.957	0.850	0.862	0.513	1.104	0.760
Armenian	1.008	0.979	1.366	0.956	1.321	0.974	0.669	0.698	0.743	0.587	0.901	0.663
Welsh	1.309	0.985	1.174	0.961	1.642	0.956	1.015	0.786	0.770	0.489	1.313	0.639
Basque	0.952	0.991	1.158	0.956	1.413	0.970	0.762	0.764	0.661	0.720	1.155	0.667
Hebrew	0.881	0.949	1.556	0.940	1.142	0.938	0.503	0.658	0.745	0.447	0.676	0.617
Cebuano	1.460	0.974	1.856	0.921	1.217	0.965	0.654	0.373	1.168	0.613	0.571	0.239
Tagalog	1.278	0.979	1.619	0.918	1.187	0.943	0.706	0.682	1.047	0.583	0.683	0.618
Chichewa	0.942	0.979	1.523	0.928	1.155	0.954	0.529	0.626	0.819	0.551	0.648	0.095
Luganda	0.916	0.972	1.446	0.941	1.196	0.939	0.569	0.692	0.761	0.557	0.751	0.587
Somali	1.090	0.979	1.276	0.958	1.295	0.970	0.835	0.813	0.848	0.695	0.973	0.665
Haitian	1.518	0.971	2.085	0.912	1.730	0.951	0.603	0.363	0.864	0.411	0.702	0.012
Nahuatl	1.205	0.956	1.205	0.956	1.398	0.938	0.544	0.448	0.857	0.559	0.642	0.067
Range	0.881 1.518	0.949 0.994	1.123 2.085	0.899 0.969	1.107 1.730	0.938 0.974	0.503 1.015	0.152 0.883	0.561 1.168	0.411 0.720	0.571 1.356	0.012 0.760

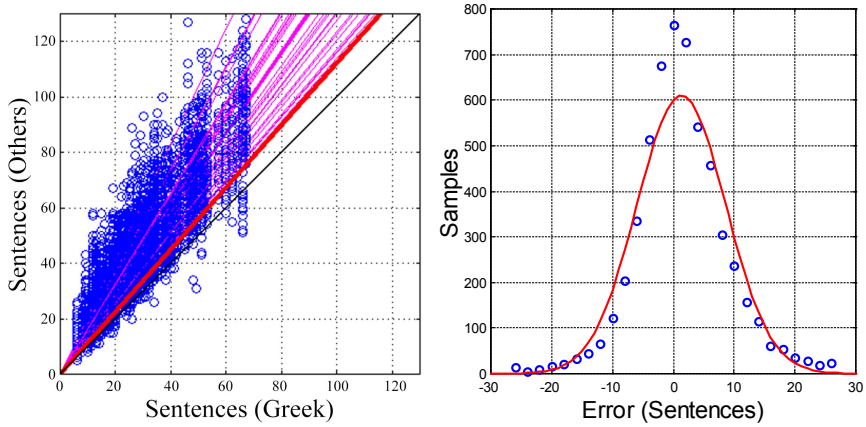


Figure 2. **Left:** Scatterplot between n_S in Greek and n_S in the other translations listed in Table 1, together with the regression lines (1). The black line is the line $y = x$. The red line is the regression line between Latin and Greek. **Right:** Histogram of the difference ("error") between the actual number of words in a given translation and the number of words in that translation calculated from the regression line, for a given Greek value.

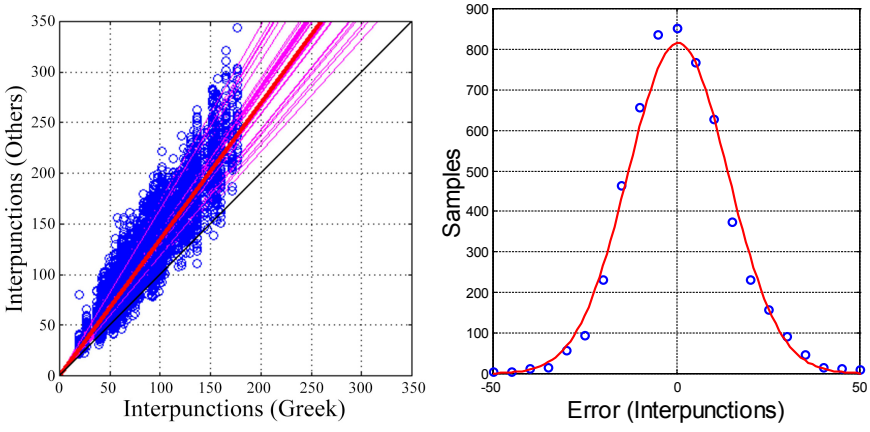


Figure 3. **Left:** Scatterplot between n_I in Greek and n_I in the other translations listed in Table 1, together with the regression lines (1). The black line is the line $y = x$. The red line is the regression line between Latin and Greek. **Right:** Histogram of the difference ("error") between the actual number of words in a given translation and the number of words in that translation calculated from the regression line, for a given Greek value.

A larger spread can be noticed in the deep-language variables P_F, I_P and M_F, Table 3 and Figures 4-6. The slopes and correlation coefficients of these variables clearly underline the fact that the distribution of interpunctions, within a chapter, introduced in any text for better conveying the meaning to readers, can be quite different from translation to translation. Compared to n_W, n_S and n_L, the multiplicative bias increases for all languages, with very few exceptions (e.g. Esperanto and Welsh in the variable P_F), and the correlation coefficients become smaller.

Now, to study the chaotic data reported in Tables 1-3, it is very useful to consider a translated text as the output of a communication channel fed by the original text. The characteristics of this channel (one for each stochastic variable) can give us more insight into the mathematical/statistical deep structure of al-

phabetical (and possibly human) languages. Before doing so, in the next section we define a useful parameter, namely the noise-to-signal power ratio of a real translation channel compared to the ideal channel.

4. Noise-to-Signal Power Ratio and Its Universal Geometrical Representation

We characterize any translation and its linguistic stochastic variables as a complex communication channel, made of parallel channels—one for each variable—affected by "noise". The input language is the "signal", the output language is a "replica" of the input language, but largely perturbed by noise. From the point of view of the output language this noise is, of course, indispensable for conveying the meaning to readers of the output language. To study these channels, we define a suitable noise-to-signal power ratio and use a geometrical representation borrowed from author's design of deep-space radio links [31], also applied in [32]. This geometrical representation is universal.

Two variables y and x, linked by a regression line $y = mx$, where m is the slope of the line, are perfectly correlated if the correlation coefficient $r = 1$, and are not biased if $m = 1$, in other words, if the regression line is $y = x$ (45˚ line, $m = 1$) and all y-values lie on the line ($r = 1$). If these conditions are not met, we consider the variance of the difference between the regression line values ($m \neq 1$) and the ideal line $y = x$ values, at given x-values, as the "regression noise" power N_m, and the variance of the difference between the values not lying on the line and the regression line $y = mx$, ($r \neq 1$), as the "correlation noise" power N_r.

Let us apply these concepts to language translation. Defined the variance s_x^2 of language x and s_y^2 of language y, the difference y - x between the regression line of the real translation channel and that of the ideal channel is given by $(m-1)x$, therefore the variance (or power) of the regression noise is given by:

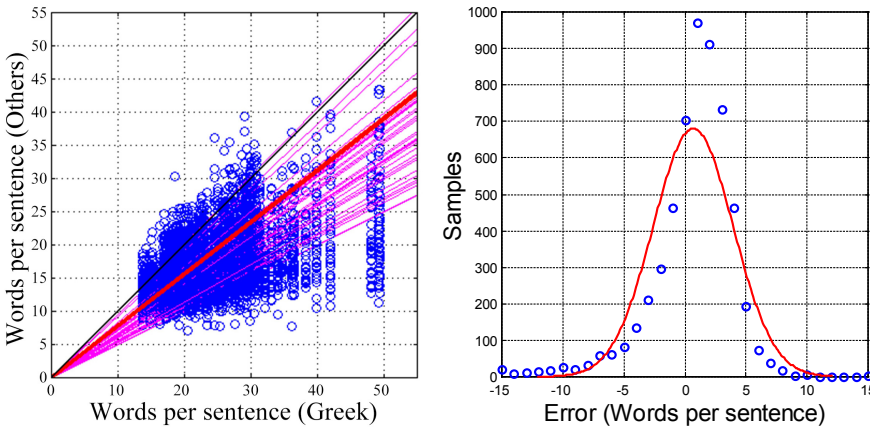


Figure 4. Left: Scatterplot between P_F in Greek and P_F in the other languages listed in Table 1, together with the regression lines (1). The black line is the line $y = x$. The red line is the regression line between Latin and Greek. **Right:** Histogram of the difference ("error") between the actual number of words in a given translation and the number of words in that translation calculated from the regression line, for a given Greek value.

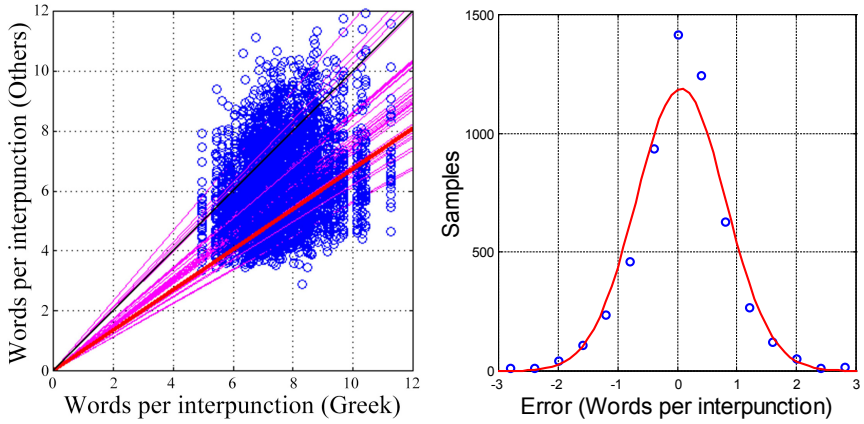


Figure 5. **Left:** Scatterplot between I_P in Greek and I_F in the other translations listed in Table 1, together with the regression lines (1). The black line is the line $y = x$. The red line is the regression line between Latin and Greek. **Right:** Histogram of the difference ("error") between the actual number of words in a given translation and the number of words in that translation calculated from the regression line, for a given Greek value.

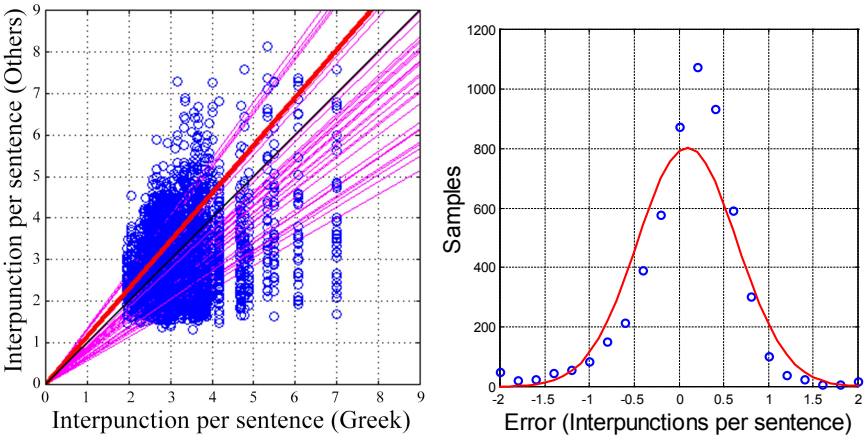


Figure 6. **Left:** Scatterplot between M_F in Greek and M_F in the other translations listed in Table 1, together with the regression lines (1). The black line is the line $y = x$. The red line is the regression line between Latin and Greek. **Right:** Histogram of the difference ("error") between the actual number of words in a given translation and the number of words in that translation calculated from the regression line, for a given Greek value.

$$N_m = (m-1)^2 s_x^2 \tag{2}$$

Then, the regression noise-to-signal power ratio, R_m, is given by:

$$R_m = \frac{N_m}{s_x^2} = (m-1)^2 \tag{3}$$

Notice that in (3) what counts is the absolute difference $|m-1|$ because R_m is an even function (parabola) around $m = 1$.

According to the theory of regression lines [30], the fraction of the variance s_y^2 due to the y-values not belonging to the line (correlation noise power, N_r) is given by:

$$N_r = (1-r^2) s_y^2 \tag{4}$$

This noise power is correlated with the slope m, because the fraction of the variance s_y^2 due to the regression line $y = mx$, namely $r^2 s_y^2$, is related to m according to the following relationship [10]:

$$r^2 s_y^2 = m^2 s_x^2 \tag{5}$$

Therefore, the correlation noise-to-signal power ratio, R_r, is given by:

$$R_r = \frac{N_r}{s_x^2} = \frac{1-r^2}{r^2} m^2 \tag{6}$$

Now, because the two noise sources are disjoint, the total noise-to-signal power ratio of the channel is given by:

$$R = R_m + R_r \tag{7}$$

By (3) and (6), R depends only on the two parameters m and r of the regression line (**Table 3**), given by:

$$R = (m-1)^2 + \frac{1-r^2}{r^2} m^2 \tag{8}$$

For each couple of the same variable, in Greek and in a translation, we can represent Equation (8) graphically by considering the variables (not to be confused with translations):

$$X = \sqrt{R_m} \tag{9a}$$

$$Y = \sqrt{R_r} \tag{9b}$$

By setting $R = R_o$, being R_o a constant, X and Y trace a circle with radius $\sqrt{R_o} = \sqrt{R_m + R_r}$ in the first Cartesian quadrant. All points inside the circle correspond to $R < R_o$; the origin of the axes corresponds to $R = 0$ of the ideal channel, $m = 1$ and $r = 1$. The reciprocal of R is the signal-to-noise power ratio $\rho = 1/R$, which becomes infinite at the origin and decreases as the radius of the circle increases.

As discussed in [31], among other features not of interest here, adopting the noise-to-signal power ratio instead of the more common signal-to-noise power ratio allows this graphical representation, which immediately shows how R_r and R_m, through their square roots, contribute to the total R, and which of the two pushes the translation away from the ideal self-translation.

In conclusion, the comparison between any couple of corresponding variables can be studied as a "communication channel" in which the input signal is the Greek text variable and the output signal is the translation variable. Compared to the ideal channel, the actual channel is noisy, always characterized by $R > 0$. Of course, as already noted, this indispensable "noise" is what actually makes the translation intelligible to the intended readers of the translated texts. In the next section we study these communication channels.

5. Linguistic Communication Channels

We compare, for each chapter, the numbers of words, sentences, interpunctions,

and the so-called deep-language variables P_F, I_F, M_F, of the original Greek texts to those of another language. The values of the slope m of the linear model (1) and the correlation r for all variables and translations can be read in Table 3. From these data we can calculate $X = \sqrt{R_m}$, $Y = \sqrt{R_r}$ and the noise-to-signal power ratio.

Let us first consider the words channel n_W. Figure 7 shows the results obtained according to the geometrical representation discussed in Section 4. The closer the point is to the origin, the less noisy the channel, therefore implying a communication channel is closer to the ideal channel. Latin, Basque, Russian and Croatian are the least noisy languages (the black circles will be discussed in Section 6). All other languages values lie approximately along the regression line:

$$Y = 0.477X + 0.157 \tag{10}$$

A regression line $Y = aX + b$ with $a > 0$, as Equation (10), is due to languages with $m > 1$, while a regression line with $a < 0$ is due to languages with m < 1.

From Equation (10) it turns out that, even though some translations can be practically unbiased ($m \approx 1$), as is the case of Slovak, they can never be perfectly correlated with the Greek texts, *i.e.*, their correlation coefficient can never approach 1. In fact, when $m = 1$, *i.e.* $X = 0$, from Equation (10) we get $Y = 0.157$ and, by setting $m = 1$ in Equation (6), we can calculate the corresponding "irreducible" (minimum) correlation coefficient:

$$r_{m=1} = 1/\sqrt{1+b} = 0.930 \tag{11}$$

This value has to be compared with the minimum value 0.949 of Hebrew (Table 3).

In conclusion, even though the channel is very close to being ideal for the slope ($m \approx 1$, no bias on the average, very small regression noise), it can never be ideal for the correlation coefficient, therefore there is always some significant correlation noise around the 45˚ line. Notice that there is no clear trend for the various language families, except for the Balto-Slavic family, which minimizes the regression noise X, because $m \approx 1$, therefore these translations are grouped towards the Y-axis. The noisiest languages are Norwegian, Cebuano and Haitian.

Let us consider the sentences channel n_S, whose results are shown in Figure 8. Now, both $X = \sqrt{R_m}$ and $Y = \sqrt{R_r}$ are further away from the origin than those of the words channel, therefore the noise-to-signal power ratio is greater than that of the words channel. Latin is, again, the least noisy language, together with Croatian and Basque. Moreover, as already noticed, the number of sentences tend to be larger than in Greek, therefore $m > 1$. The noisiest language is Haitian because of the extreme values $m = 2.085$ and $r = 0.912$. The regression line drawn in Figure 8 is given by $Y = 0.751X + 0.178$, therefore the irreducible correlation coefficient is $r_{m=1} = 0.921$ approximately the same of the words channel. In other words, if there were no multiplicative bias ($m = 1$), the spread

of words and sentences around the regression lines, Table 3, would be very similar. Now, because characters and words are very much correlated ($r > 0.990$ for all languages, not shown but verified, just like for Italian literature [28]), this observation applies also to the characters channel.

Let us consider the interpunctions channel n_I, whose results are also shown in Figure 8. This channel is noisier than n_W and n_S channels. Swedish is the least noisy language, Haitian the noisiest. Each language, in fact, introduces a very different distribution of interpunctions in a chapter, both in type (full-stops, question marks, exclamation marks, commas, colons, semicolons) and quantity, therefore changing the length of sentences, word intervals, and interpunctions per sentence. The regression line drawn in Figure 8 is given by $Y = 0.397X + 0.256$, therefore the irreducible correlation coefficient (11) is $r_{m=1} = 0.892$, the lowest of the three channels examined so far.

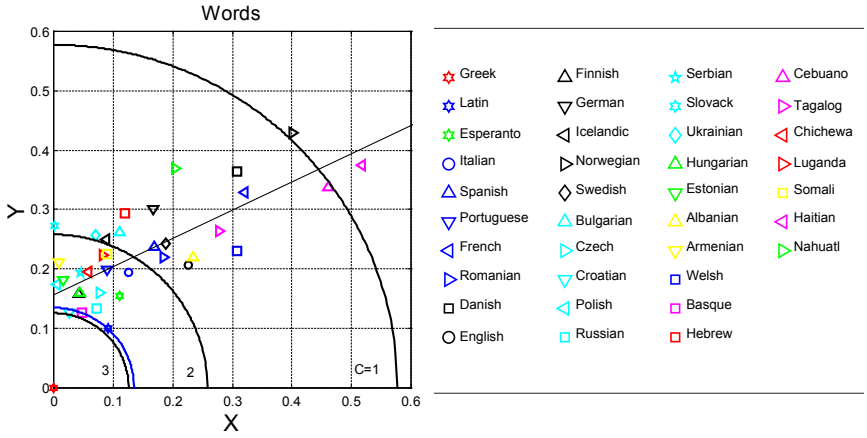


Figure 7. Scatterplot between $X = \sqrt{R_m}$ and $Y = \sqrt{R_r}$ in the words n_W Channel. The origin represents the ideal channel. The black arcs of circles give contours of equal channel capacity C (Section 6).

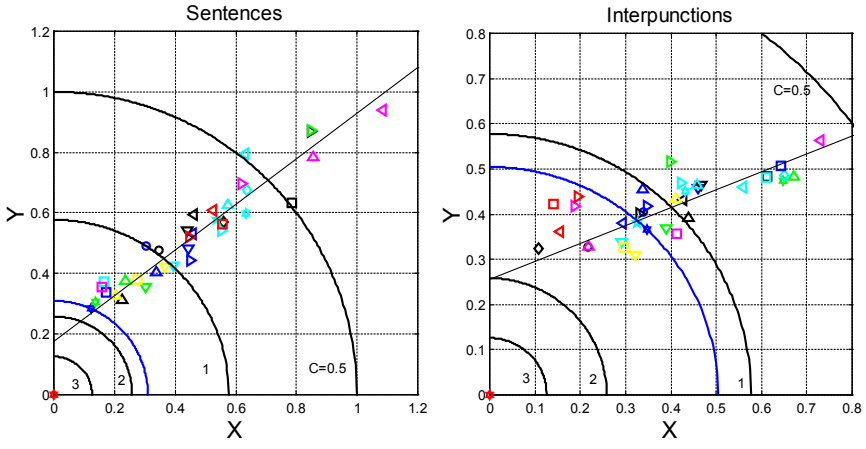


Figure 8. Scatterplot between $X = \sqrt{R_m}$ and $Y = \sqrt{R_r}$. The origin represents the ideal channel. **Left:** n_S channel. **Right:** n_I channel. The black arcs of circles give contours of equal channel capacity C (Section 6).

Let us consider the number of words per sentence, P_F, a deep-language variable. Figure 9 shows the results obtained. This channel has a large correlation noise, as we can see from the range of Y, a consequence of the very low correlation coefficients (Table 3). The least noisy language, again, is Latin, the noisiest is Norwegian.

The results of the channel concerning the number of words per interpunction, *i.e.* the word interval I_P, are also shown in Figure 9. The least noisy languages are Basque, Latin, Estonian and Croatian, the noisiest is Haitian. In general, the I_P channel is less noisy than the P_F channel. It seems that I_P cannot be set as much independently from Greek as P_F seems it can be. A likely explanation is that the word interval is empirically correlated with the short-term memory capacity, and this capacity not only is limited according to the 7 ± 2 Miller's law [27], but it cannot change so much in humans, regardless, of course, of the language used, therefore it varies less from language to language. This is not the case for P_F, a variable more linked to the output language, or translation style and intended readers through a readability index (see Section 8), than to human short-term memory capacity.

The results of the channel concerning the number of interpunctions per sentence, M_F, are also shown in Figure 9. The least noisy language is again Croatian, the noisiest is again Haitian, with $Y \approx 60$ (due to the very low correlation coefficient 0.012, practically zero) and $X \approx 0.3$, not shown because much out of scale. Notice that I_P and M_F channels are quite similar for most languages.

Compared to n_W, n_S and n_S channels, the deep-language variables channels are the noisiest. The reason seems to be, again, the different distribution of interpunctions. For these channels we have not drawn regression lines because the correlation coefficient is small.

Let us summarize the main results of this section. The channels studied are differently affected by the translation noise. The most accurate channel is the word channel n_W, a finding that seems reasonable. Humans seem to express a given meaning with a number of words—*i.e.* finite strings of abstract signs (characters)—which cannot vary so much even if some languages (Hebrew, Welsh, Basque etc.) do not share, according to scholars, a common ancestor with most other languages. This result seems to be something basic to human processing capabilities.

The number of sentences and their length in words, *i.e.* P_F, can be treated more freely. We know that P_F affects readability indices very much, as shown for Italian [28], therefore, this variable tends to be better matched to the intended readers, with specific reading ability, not to the original Greek readers of the Roman Empire.

Finally, we observe that, independently of the different channels, the correlation noise is always larger than the regression noise, therefore indicating that every translation tries as much as possible not to be biased, but it cannot avoid being decorrelated, with correlation coefficients which approximately decrease from words, to sentences, to interpunctions and down to the deep-language variables.

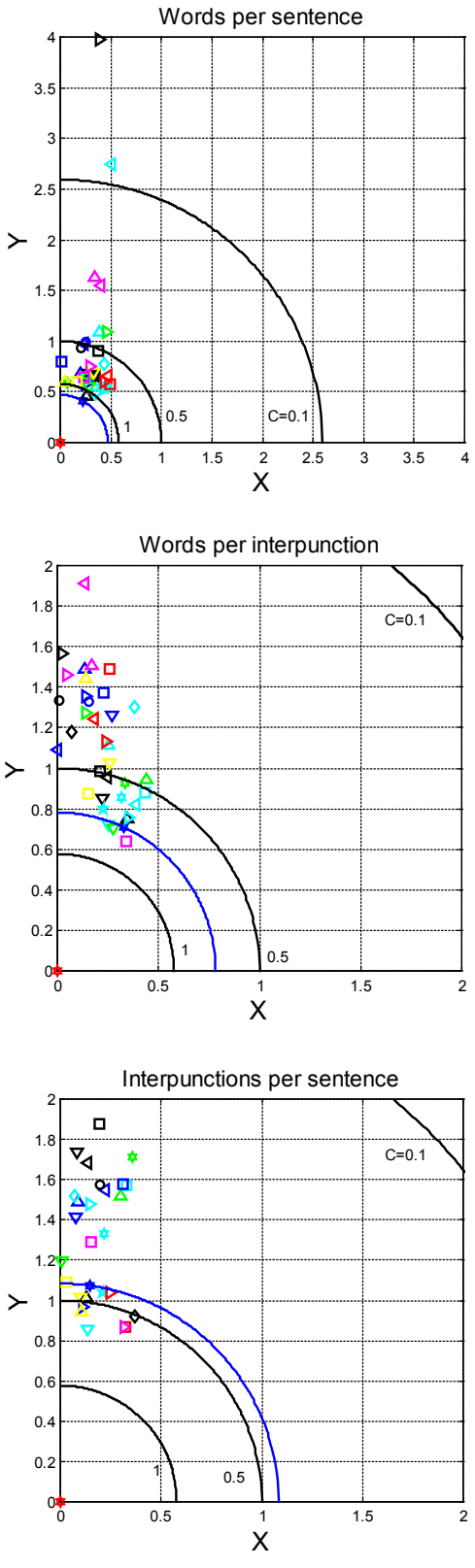


Figure 9. Scatterplot between $X=\sqrt{R_m}$ and $Y=\sqrt{R_r}$. The origin represents the ideal channel. **Upper:** P_F channel. **Middle:** I_P channel. **Lower:** M_F channel. The black arcs of circles give contours of equal channel capacity C (Section 6).

Besides the noise-to-signal power ratio, communication channels can be also characterized by the channel capacity, as we discuss in the next section.

6. Channel Capacity

The noise-to-signal power ratio and its universal geometrical representation is not the only interesting way for studying noisy channels. Noisy channels can be also characterized by a single variable, namely the channel capacity or mutual information defined by Shannon [26], between the stochastic variables x (input) and y (output), see also [33]. In the following subsections, firstly we recall the channel capacity of communication theory and define what we mean by "symbol"; secondly, we assess, for the first time, the size of channel capacity obtainable with linguistic variables.

6.1. Channel Capacity According to Communication Theory

According to Shannon [26], under some assumptions, the capacity (bits per symbol) of the channel $X \to Y$ is related directly to the channel signal-to-noise power ratio $1/R$, according to:

$$C = 0.5 \times \log_2 \left(1 + 1/R\right) \tag{12}$$

In our analysis the term "symbol" is defined according to the linguistic variable under study. For example, in the words channel the "symbol" is defined as the number of words per chapter, therefore, the actual values n_W of input and output chapters. For example, in Matthew, Chapter 5, the input symbol (Greek) is 823, while the output symbol is 1006 in English, 932 in Italian and 765 in Russian. Therefore, the magnitude of additive noise is 1006 − 823 = 183 in English, 932 − 823 = 109 in Italian and 765 − 823 = −58 in Russian. This noise can be relatively large as it peaks at 22.2% of the input value in the English translation. The signal-to-noise power ratio of this sample is, therefore, $(823/183)^2 = 20.2$ in English and $(823/58)^2 = 201.3$ in Russian, synthetically underling that the Russian translation is closer to Greek than the English translation.

In other words, we do not consider the classical information content of texts according to communication/information theory, which, to a first approximation, is measured by the entropy of letters [34], a concept applicable to machine translation but not to human information processing, which is based on words, sentences and interpunctions distribution. Indeed, the short-term memory responds to words not to bits, therefore the use of entropy can be highly misleading in estimating the characteristics of the linguistic channels defined in the present paper (Appendix B).

For a constant $R = R_o$, Equation (12) gives the minimum channel capacity if the noise is Gaussian. If the noise is not Gaussian, the actual channel capacity is larger than (12) [26].

Of the two noise sources defined in Section 4, the correlation noise and the regression noise, the latter is deterministic (it could be cancelled by dividing the

variables of the output language by the corresponding m, if known), but the first can approximately be modelled as Gaussian, Figures 1-6. Therefore, if we assume that both sources of noise are Gaussian, then the channel capacity calculated with Equation (12) is pessimistic. In any case, this is not of concern here because Equation (12) can be used for comparing different translations.

We have already shown contours of constant capacity C (given, of course, by constant $R = R_o$) in Figures 7-9, namely the black arcs of circles. In the origin of the Cartesian coordinates $R = 0$, therefore $\rho = \infty$ and $C = \infty$. This last result, valid for the continuous channel assumed in Equation (12), merely means that the channel does not impose any limit to the output information, therefore in this case the mutual information coincides with the input self-information of the Greek texts.

Of the channels studied in Section 5, the words channel n_W has the largest channel capacity for most translations. Figure 10 shows the scatterplot between the capacities of n_W and n_S channels. We notice that the two channels are quite correlated; for Welsh the two capacities are even practically identical. Figure 10 shows also the scatterplot between the capacities of n_W and n_I channels. The two capacities are practically uncorrelated. In Appendix C we report the scatterplots of the capacities of words channel and sentences channel with the deep-language channels capacities. In all cases, we notice a poor correlation, except partially for the P_F channel, therefore evidencing, again, the fact that every translation has its own pattern of interpunctions within a chapter, which determines P_F, I_P and M_F.

Some interesting observations can be done on the mixed scatterplots shown in Figure 11 between I_P and n_W, n_S and I_P channels capacities. The correlation between these variables is evident: as I_P increases, thus loading more reader's short-term memory, the channel capacities decrease. In other words, by decreasing this important deep-language variable, I_P, channels tend to be closer to the ideal channels of words, sentences and I_P itself.

Differently of the word interval I_P, the number of words per sentence P_F is quite correlated only with its channel capacity, Figure 12. As P_F approaches the Greek value (23.07, Table 2), the channel capacity increases. This different behavior compared to Figure 11 where, as I_P approaches the Greek value 7.47, I_P channel capacity decreases, underlines that I_P seems to be more related to how human brain processes texts (short-term memory), regardless of the particular language. In other words, translations do not follow the high Greek I_P. On the contrary, P_F is more related and matched to the intended readers through the readability index, which does not consider I_P [28].

In the next subsection we discuss how large is the capacity of linguistic channels.

6.2. Channel Capacity Size

Two questions arise: 1) Are the channel capacities large? 2) How can we assess how large they are? Let us start with studying the sensitivity of the channel capacity to the parameters m and r. Figure 13 shows a universal chart, drawn from

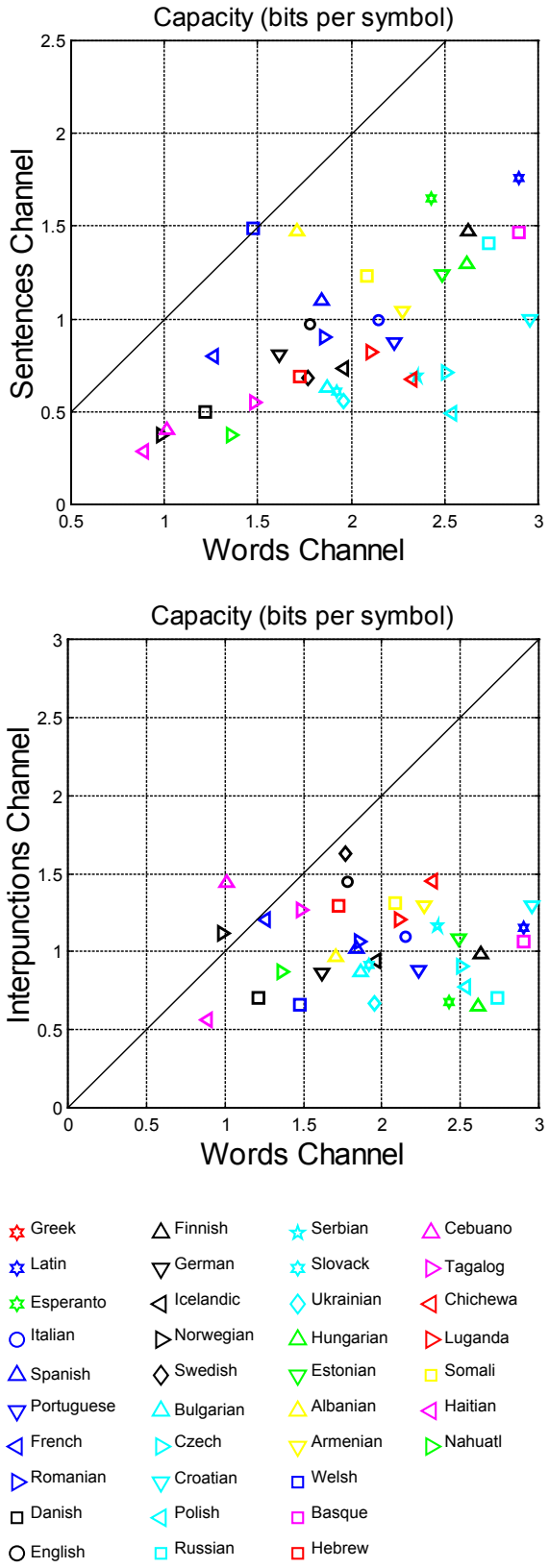


Figure 10. Upper: Scatterplot between the capacities of n_W and n_S channels. **Middle:** Scatterplot between the capacities of n_W and n_I channels. **Lower:** symbols caption.

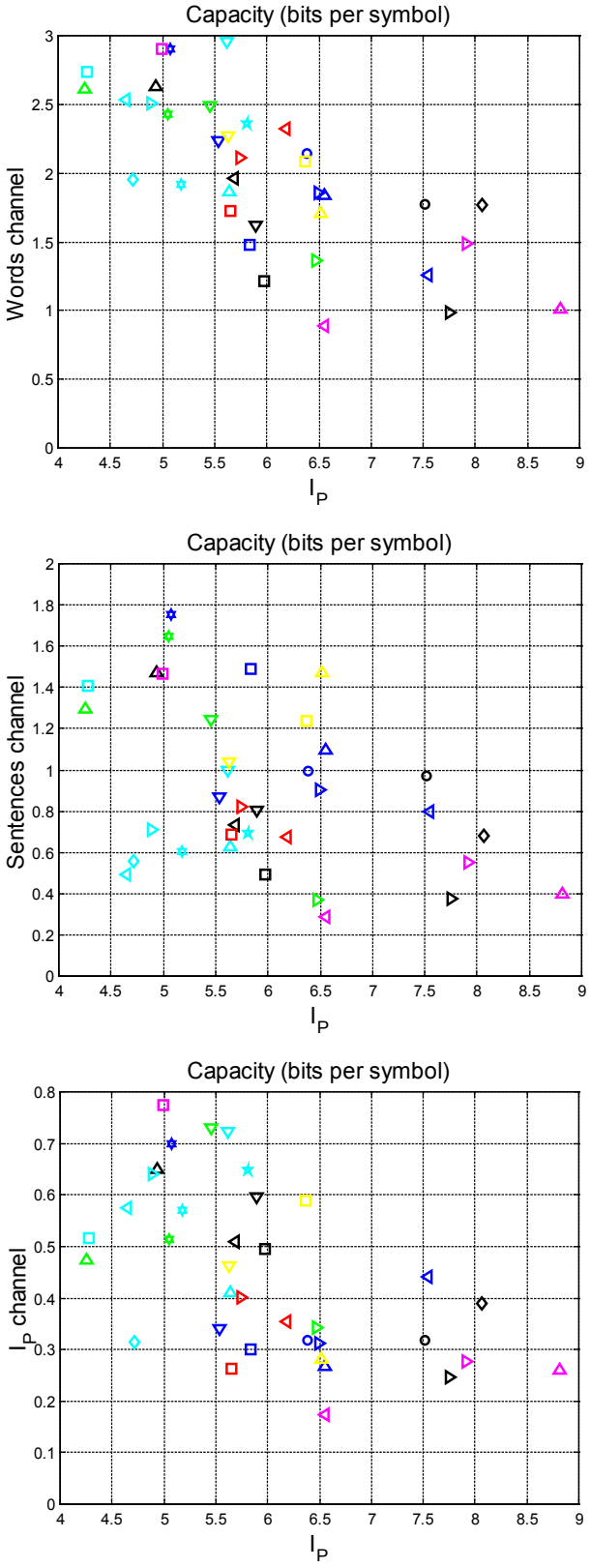


Figure 11. Upper: Scatterplot between I_P and the capacity of n_W channel. **Middle:** Scatterplot between I_P and the capacity of n_S channel. **Lower:** Scatterplot between I_P and the capacity of I_P channel.

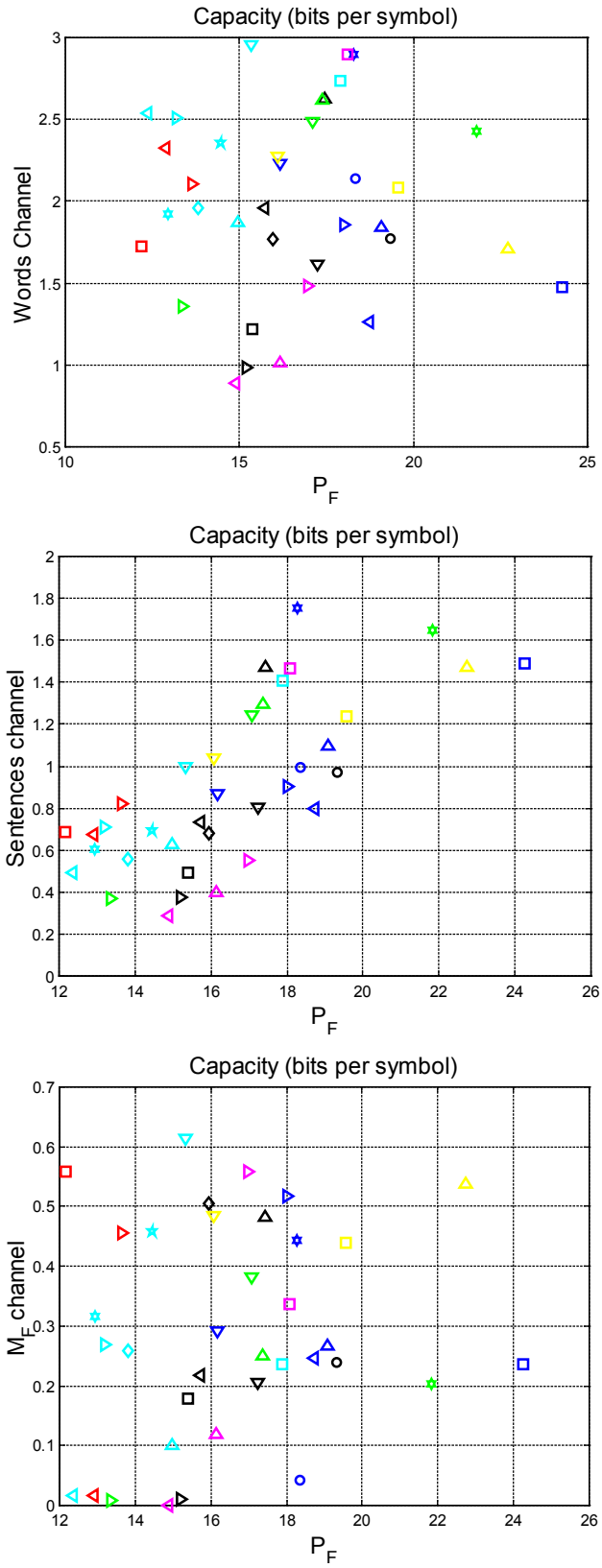


Figure 12. Upper: Scatterplot between P_F and the capacity of n_W channel. **Middle:** Scatterplot between P_F and the capacity of n_S channel. **Lower:** Scatterplot between P_F and the capacity of M_F channel.

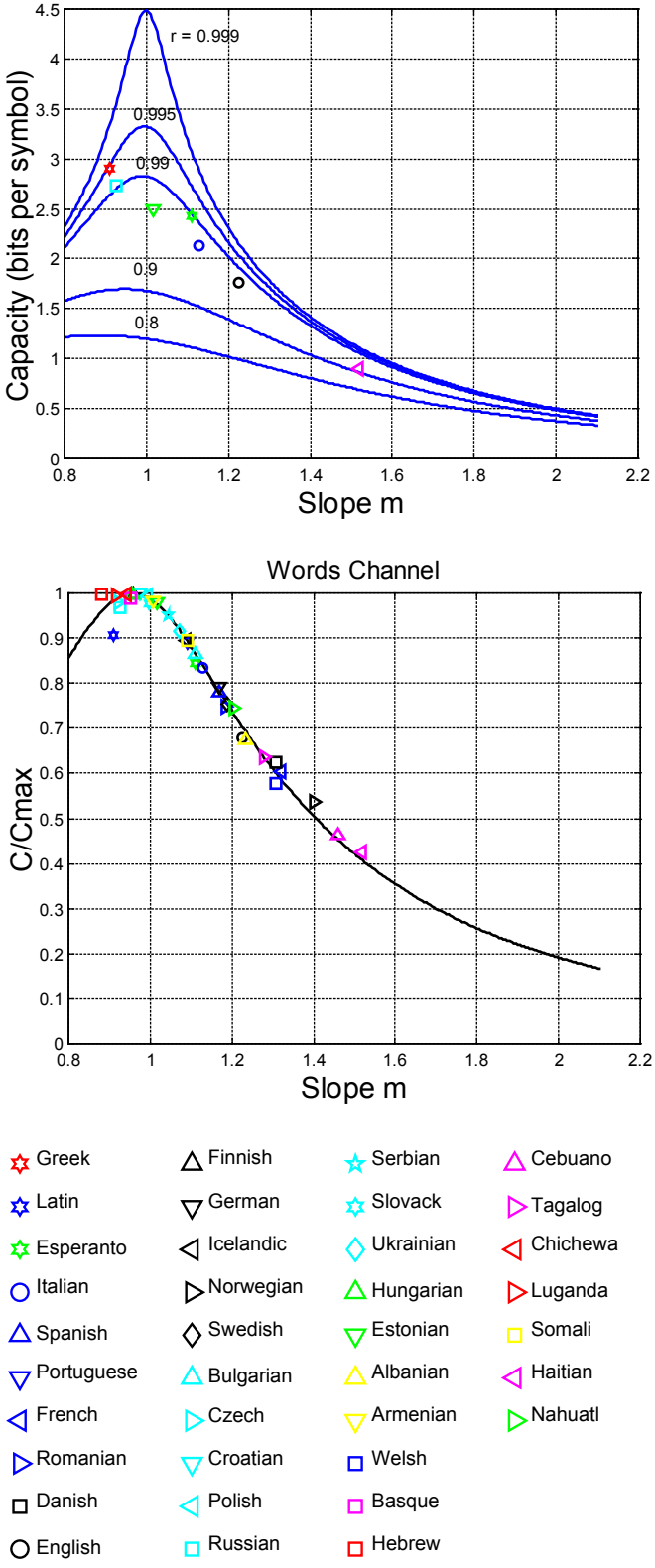


Figure 13. Upper: Universal chart describing the relationship between the channel capacity C and the slope m, as a function of the correlation coefficient r. For illustration, the values of the n_W channel capacity of some translations are also shown. **Middle:** C/C_{max} of the n_S channel. **Lower:** symbols caption.

Equations (12) and (8), which describes the relationship between the channel capacity C and the slope m, as a function of the correlation coefficient r. For illustration, we have also reported the values of the words channel capacity of some translations.

The maxima of C are found from Equation (12) when $R = R_{\min}$, which occurs if:

$$m_{C\max} = r^2 \tag{13}$$

Therefore, from (8) it follows

$$R_{\min} = 1 - r^2 \tag{14}$$

Consequently, from (12) we get:

$$C_{\max} = 0.5 \times \log_2\left[1 + \frac{1}{1-r^2}\right] \tag{15}$$

Because of (15), in Figure 13 we can notice a very sharp increase only for very high correlation coefficients. In actual translations, however, the capacity can be significantly large, not too far from the maximum value obtainable from Equation (15). In fact, defined the normalized capacity $C/C_{\max}$, Figure 13, Figure 14 show how $C/C_{\max}$ varies. Notice that $C/C_{\max}$ practically follows the same mathematical function, regardless of the channel (words or sentences) when the correlation coefficient r is about the same for all languages (Table 3). The same result is also found for the interpunctions channel (not shown for brevity). For P_F and I_P channels (Figure 14) no regularity emerges because of poor correlation coefficients, another sign that these deep-language variables depend more profoundly on the particular translation, not on the language. The M_F channel follows the same trend (not shown).

In conclusions, the capacity of n_W, n_S and n_I channels follow very closely the universal chart because of similar high correlation coefficients; on the contrary, the capacity of P_F and I_P channels is more spread because their correlation coefficients greatly varies from translation to translation.

7. Word Interval and Short-Term Memory

As studied and discussed in [28], the number of words per interpunctions, namely the word interval I_P, varies in the same range of the short-term memory capacity—given by the 7 ± 2 Miller's law [27], a range where 95% of all occurrences are found—and is very likely related to it because interpunctions organize small portions of more complex arguments in short chunks of text. Moreover, drawn I_P against the number of words per sentence P_F, I_P tends to saturate to a horizontal asymptote as P_F increases. In other words, even if sentences get longer, I_P cannot get larger than about the upper limit of Millers' law (namely 9), because of the constraints imposed by the short-term memory capacity of readers.

Empirically (best-fit) the average value of I_P is related to the average value of P_F according to the relationship [28]:

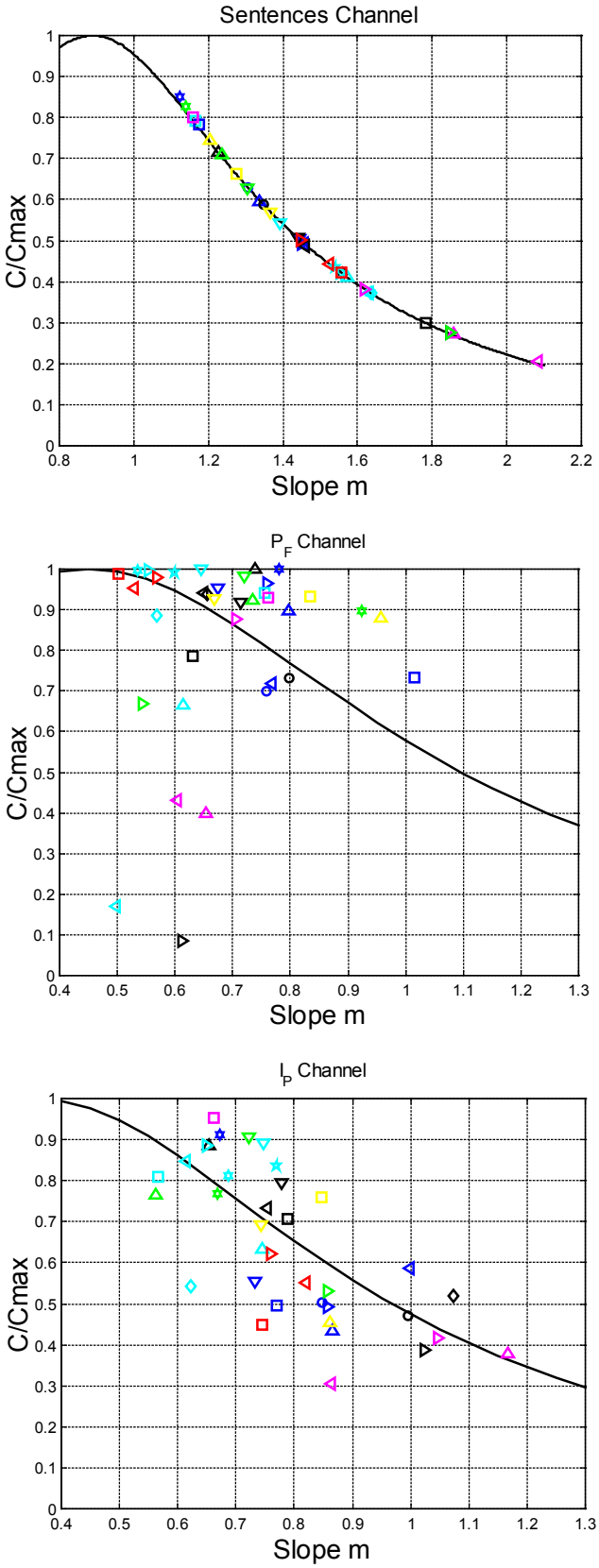


Figure 14. Upper: C/C_{max} of the n_S channel. **Middle:** C/C_{max} of the P_F channel. **Lower:** C/C_{max} of the I_P channel.

$$I_P = \left(I_{P\infty} - 1\right) \times \left[1 - e^{-\frac{(P_F - 1)}{(P_{Fo} - 1)}}\right] + 1 \tag{16}$$

where $I_{P\infty}$ gives the horizontal asymptote, and P_{Fo} gives the value of P_F at which the exponential falls at $1/e$ of its maximum value. We apply Equation (16) to the NT translations. Because both I_P and P_F depend on the translation, we find different constants in Equation (16), listed in **Table 4**, together with data concerning readability index discussed in Section 8.

Figure 15 shows the scatterplot concerning Greek, Latin and Hebrew. As for the Italian Literature (see **Figure 16** of [28]), I_P spreads in Miller's range. Not surprisingly, the ancient readers of these texts had the same short-term memory capacity of modern readers, *i.e.* they followed Miller's 7 ± 2 law. This finding is confirmed by the results concerning modern languages for which, however, the spread within Miller's range can be different from translation to translation. Some translations tend to use shorter values of I_P, as Latin and Hebrew (**Figure 15**), therefore loading less reader's short-term memory than other translations do, e.g. Italian, French and English (see asymptote values $I_{P\infty}$ in **Table 4**). In Appendix D we show more graphical examples.

Figure 16 shows all best-fit models of **Table 4** and also the best-fit for Greek, with ±1 standard deviation calculated from the models of **Table 4**. We see that Miller's lower bound $I_p = 5$ corresponds to $P_F = 10$, therefore this value sets approximately a lower bound to the average length of sentences, a result generally valid for all languages considered.

In conclusion, each translation tends to address readers with different reading abilities because small I_P values are better matched to readers with small short-term memory capacity, who, therefore, can handle only short sentences, which correlates well with a large readability index, as we show in the next section.

8. Readability Index

As discussed in [28], after an in-depth review based on many references there listed—to which we refer readers for further details—a readability formula gives an index that anyone can calculate directly and easily, so that a writer can sufficiently match text and expected readers. Its "ingredients" are understandable by anyone, because they are interwound with long-lasting writing and reading experience based on characters, words and sentences. A readability formula gives an index based on the same stochastic variables, regardless of the text considered, thus it provides an objective measurement for comparing different texts, or authors. A final objective readability formula—or software-developed methods—is very unlikely to be found or accepted by everyone. On the contrary, instead of absolute readability, readability differences can be more useful and meaningful. The classical readability formulae provide these differences easily and directly.

Table 4. Constants $I_{P\infty}$ and P_{Fo} of Equation (16) for each translation. Average and standard deviation of the readability index G for each translation. Slope m and correlation coefficient r of the regression line between G in Greek and G in the other languages.

Language	$I_{P\infty}$	P_{Fo}	G	m	r
Greek	9.27	13.61	58.44 (4.27)	1	1
Latin	6.15	10.50	62.06 (5.17)	1.058	0.889
Esperanto	6.34	14.06	59.02 (3.77)	1.008	0.848
French	9.46	11.78	60.63 (2.78)	1.037	0.751
Italian	9.66	17.56	61.22 (3.86)	1.048	0.736
Portuguese	6.45	8.18	63.48 (3.65)	1.090	0.746
Romanian	7.28	7.76	62.02 (4.53)	1.058	0.770
Spanish	8.43	12.78	60.81 (3.80)	1.040	0.823
Danish	7.56	10.00	64.26 (3.43)	1.100	0.701
English	9.53	12.45	60.45 (3.71)	1.031	0.745
Finnish	5.65	8.31	62.76 (4.73)	0.785	0.802
German	6.75	8.36	62.37 (3.55)	1.063	0.755
Icelandic	7.19	10.16	64.13 (3.56)	1.095	0.713
Norwegian	11.21	13.05	67.04 (2.81)	1.099	0.435
Swedish	13.38	17.57	63.65 (3.56)	1.101	0.733
Bulgarian	6.65	7.90	65.06 (3.76)	1.112	0.704
Czech	5.93	7.49	68.39 (5.45)	1.168	0.722
Croatian	7.37	10.68	65.15 (4.80)	1.111	0.795
Polish	4.98	4.42	68.78 (4.54)	1.184	0.448
Russian	4.80	7.94	62.44 (5.02)	1.062	0.844
Serbian	7.11	8.49	65.90 (4.45)	1.125	0.736
Slovak	6.52	8.26	68.01 (4.66)	1.167	0.678
Ukrainian	4.76	2.73	66.60 (3.52)	1.140	0.700
Estonian	6.40	8.83	62.99 (4.64)	1.075	0.814
Hungarian	4.77	7.76	62.90 (4.25)	0.892	0.787
Albanian	8.07	13.82	58.39 (3.76)	0.996	0.781
Armenian	6.44	7.65	64.38 (5.92)	1.093	0.791
Welsh	6.25	8.65	57.31 (2.82)	1.025	0.786
Basque	5.63	8.21	61.98 (4.01)	0.691	0.801
Hebrew	6.76	6.61	69.64 (4.99)	1.192	0.682
Cebuano	11.95	11.98	63.06 (2.05)	1.079	0.400
Tagalog	8.57	6.22	62.78 (3.11)	1.071	0.592
Chichewa	12.34	19.34	68.16 (3.51)	1.166	0.659
Luganda	7.65	9.86	67.33 (4.56)	1.151	0.713
Somali	8.57	14.31	60.92 (3.94)	1.041	0.791
Haitian	9.12	11.94	64.72 (3.06)	1.109	0.519
Nahuatl	11.35	16.32	67.04 (2.81)	1.149	0.521

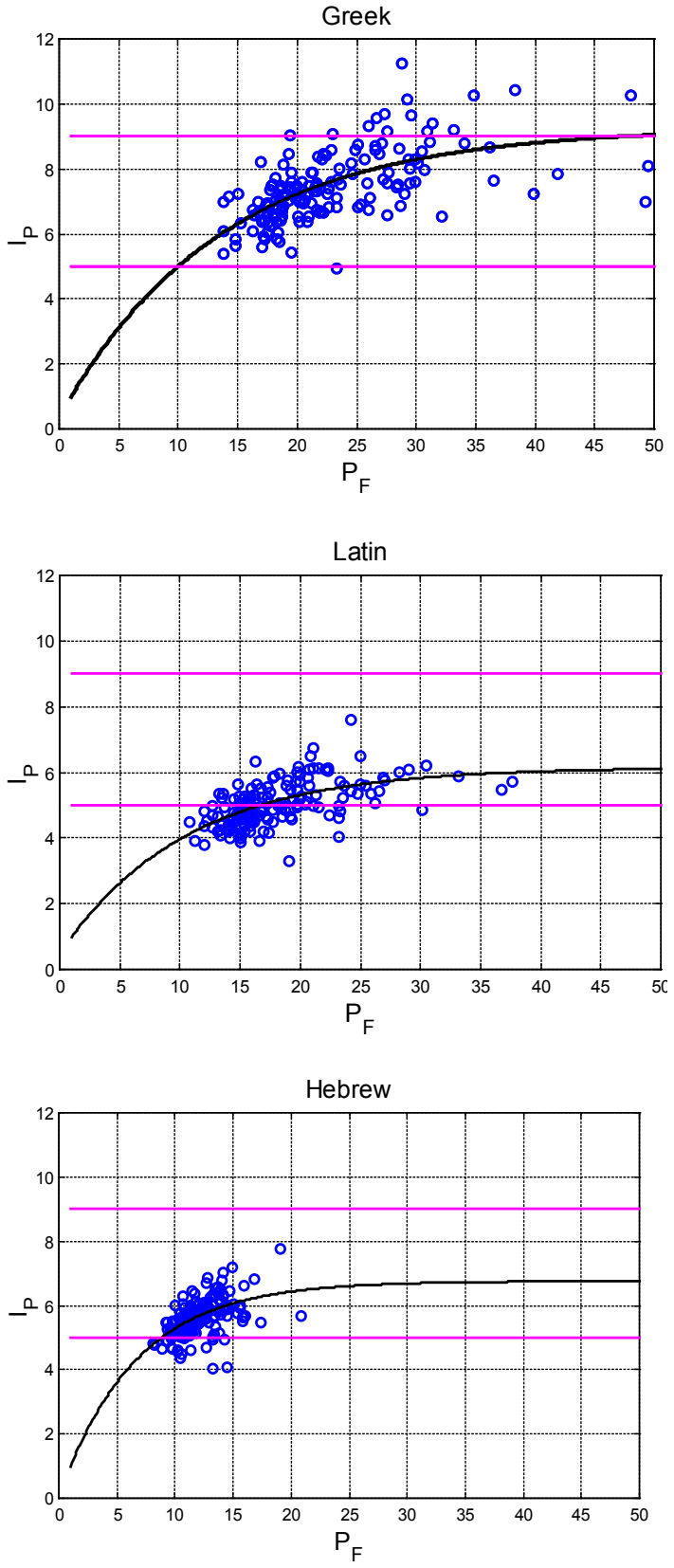


Figure 15. Upper: I_P versus P_F in Greek. **Middle:** I_P versus P_F in Latin. **Lower:** I_P versus P_F in Hebrew. Miller's bounds: magenta lines.

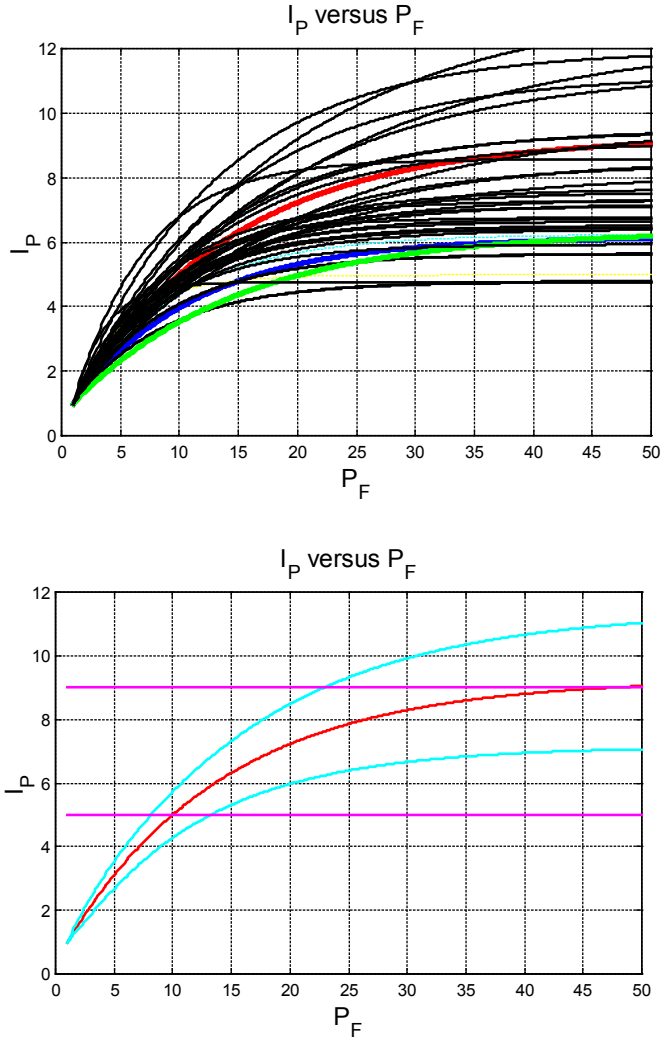


Figure 16. Upper: I_P versus P_F: best-fit from Table 4. Greek: red line; Latin: blue line; Esperanto: green line. **Lower:** I_P versus P_F: Greek, red line; ±1 standard deviation calculated from the relationships of Table 4. Miller's bounds: magenta lines.

In particular, the last observation can justify our present proposal to adopt a readability formula that can be used for comparing texts of different languages because most of them do not have a readability formula, and few adapt some formulae studied for English texts to their texts [35] [36]. The proposed formula, of course, does not exclude using other readability formulae—e.g., the large choice for English [37]—but it allows to compare, on the same ground, the readability of texts written in different languages.

For this purpose, we propose to adopt, as a calque, the readability formula used for Italian, amply studied in [28], known with the acronym GULPEASE [38], and given by:

$$G = 89 - 10 \times c/p + 300 \times f/p \tag{17a}$$

In Equation (17a) p is the total number of words in the text considered, c is the number of characters contained in the p words, f is the number of sentences

contained in the p words.

Notice that Equation (17a), as all readability formulae found in the literature, does not contain any reference to interpunctions, therefore it does not consider the very important parameter linked to the short-term memory capacity, namely the word interval I_P.

G can be interpreted as a readability index by considering the number of years of school attended in Italy's school system, as shown in Figure 17. The larger G, the more readable the text. By noting that $C_P = c/p$; $f/p = 1/P_F$, G can be written as:

$$G = 89 - 10 \times C_P + 300/P_F \tag{17b}$$

$$G = 89 - G_C + G_F \tag{17c}$$

In [28] we have shown that the term $G_c = 10 \times C_P$ (loosely referred to as the semantic term) varies very little from text to text and across centuries, while the term $G_F = 300/P_F$ (loosely referred to as the syntactic term) varies very much and, in practice, determines the readability index. We propose to use this formula also for the other languages listed in Table 1, by scaling the constant 10 of the semantic term according to the ratio between the average number of characters per word in Italian, $\langle C_{p,ITA} \rangle = 4.48$, and the average number of characters per word in another language, e.g., Greek $\langle C_{p,GRE} \rangle = 4.86$, see Table 1. The rationale for this choice is that C_P is typical of a language and, if not scaled, would bias G, without really quantifying the change in reading difficulty of readers, who are accustomed to reading in their language shorter or longer words, on the average, than those found in Italian. In other words, this scaling avoids changing G for the only reason that a language has, on the average, words shorter or longer than Italian.

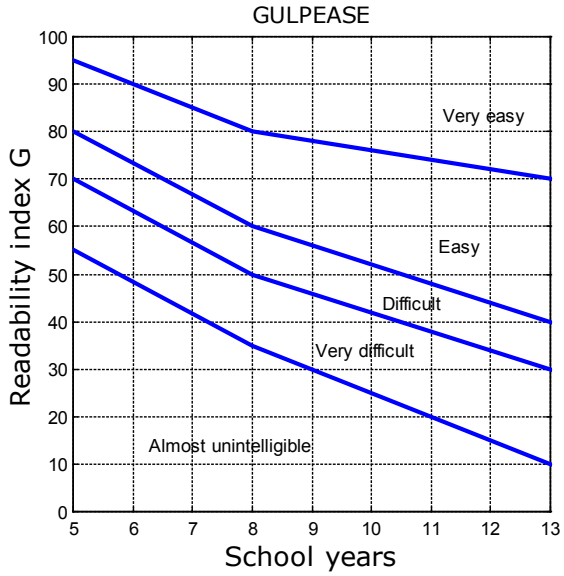


Figure 17. Readability index G versus school years in Italy, with regions of different reading difficulty.

On the other hand, we maintain the constant 300 because P_F depends significantly on reader's reading ability and short-term memory capacity [28], in other words on translator's choice. Therefore, the formula takes already care of the reader to whom the translation is addressed. Finally, notice that the constant 89 sets just the ordinate scale, therefore it has not impact on comparisons.

Therefore, the readability formula of a text written in a language with average $\langle C_P \rangle$ characters per word is given by:

$$G = 89 - 10 \times k \times C_p + 300/P_F \tag{18a}$$

with

$$k = \langle C_{p,ITA} \rangle / \langle C_P \rangle \tag{18b}$$

By using Equation (18), we force the average value of G_C to be equal to that found in Italian, namely $G_C = 10 \times 4.48$. For example (see **Table 1**), for Greek C_P is multiplied by 10 × 4.48/4.86 = 9.22, instead of 10, for Finnish (longer words) C_P is multiplied by 10 × 4.48/6.22 = 7.20 and for Haitian (shorter words) for 10 × 4.48/3.37 = 13.29.

Figure 18 shows G_C and G_F versus G, for Greek, Latin and for all languages, with some other examples shown in Appendix E. We can notice that G_F largely determines G, compared to G_C. The regression line relating G_F to G, drawn in **Figure 18**, is given by $G_F = 0.813 \times G - 32.4$. The correlation coefficient is 0.720, therefore $0.720^2 = 0.518$ is the fraction of the variance of G_F due to Equation (19). The remaining fraction 1 − 0.518 = 0.482 is due to the values scattered around the line. On the contrary, the correlation coefficient between G_C and G of the regression line $G_C = -0.187G + 56.6$ also drawn in **Figure 18**, is −0.074, practically zero, therefore confirming that G is mainly determined by G_F.

Figure 19 shows the scatterplot and the regression lines between the values of G in a translation and those in Greek, and the histogram of the difference (error) between the actual values and the regression line values. **Table 4** reports average values and standard deviations for all translations, together with the slope and correlation coefficient of the regression lines shown in **Figure 19**. As we can notice, each translation sets different readability values for their intended readers, in a large spread. In other words, as mentioned above, the number of words per sentence P_F distinguishes significantly the translations. From **Table 4** we notice that Welsh, Albanian and Greek have the lowest average G (57 - 58), making them the least readable translations, while Hebrew (69.64), followed by Polish and Czech, are the most readable translations. Now, the texts of these two extremes, to be "easy" to read according to **Figure 17**, require 8 years of equivalent Italian schooling for $G \approx 57$ and 6.5 years for $G \approx 70$. They would become "difficult", "very difficult" or even "almost unintelligible" to readers with very few years of schooling.

In conclusion, Equation (18) can be useful for comparing the readability of texts (not necessarily translations) written in different languages because of a "common ground" for interpreting them, namely **Figure 17**, which can be used as a first guide to assess readability according to the years of schooling.

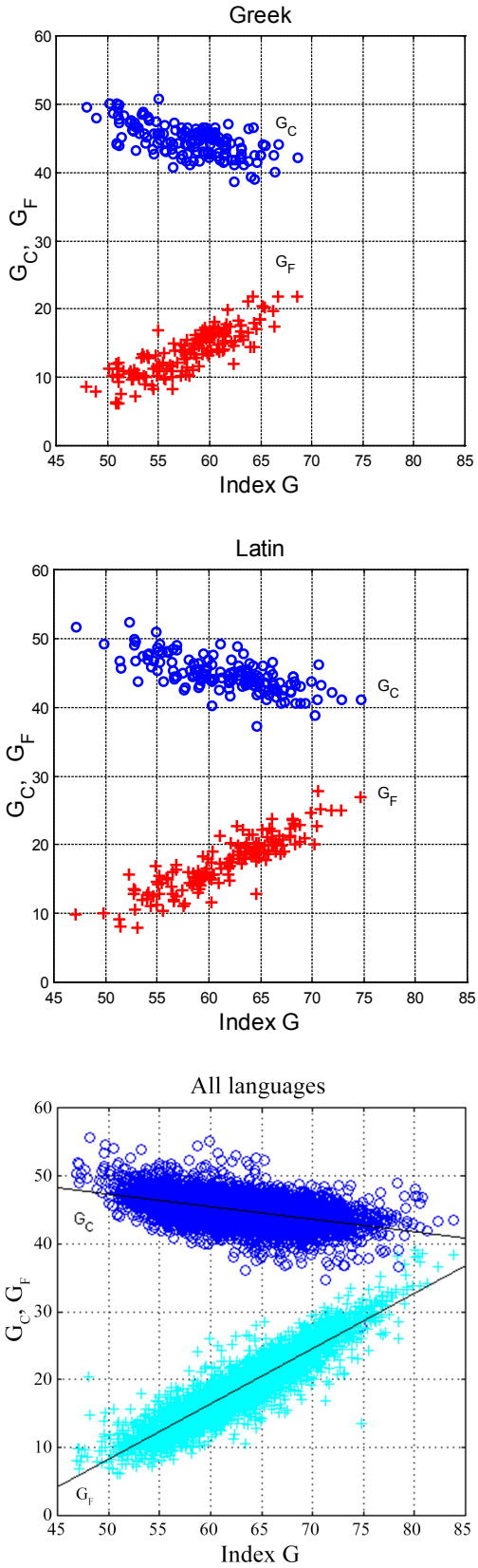


Figure 18. Upper: G_C (blue) and G_F (red) versus G in Greek. **Middle:** In Latin. **Lower:** G_C (blue) and G_F (cyan) versus G in all languages.

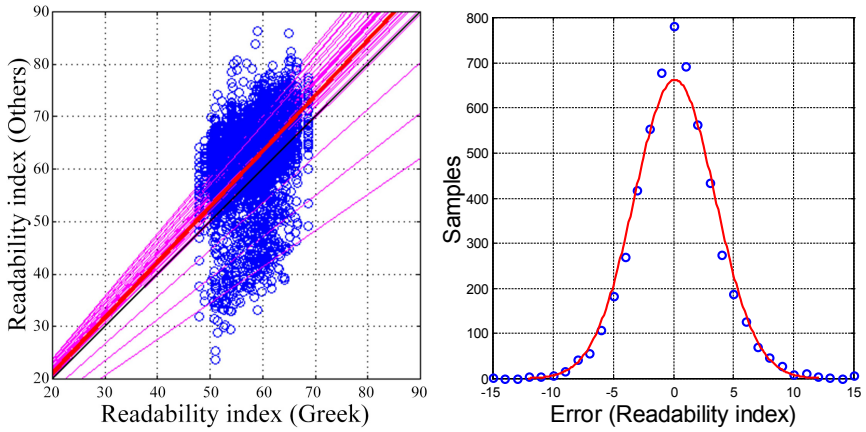


Figure 19. Left: Scatterplot between G in Greek and G in the other translations listed in Table 1, together with the regression lines (1). The black line is the line $y = x$. The red line is the regression line between Latin and Greek. **Right:** Histogram of the difference ("error") between the actual number of words in a given translation and the number of words in that translation calculated from the regression line, for a given Greek value.

9. Different NT Translations within the Same Language

If we considered different translations of the NT within the same language, do the statistics of linguistic parameters change? In other words, different versions of the NT in the same language are very similar, or do they differ from each other, maybe as much as do NT versions belonging to different languages? Indeed, for some languages there is a huge number of distinct translations: we have counted at least 60 English and 20 Spanish versions[39], which means that at least 60 different audiences have been considered in the English case and 20 in the Spanish case, which is really remarkable.

In this section, just for a very preliminary investigation, we report the average values of the most important linguistic parameters concerning 6 languages and 18 distinct versions, 3 per language, of Matthew's gospel, namely English, German, Polish, Russian, Spanish and Swedish, Table 5.

In Table 5 we notice that even the number of words and sentences can change within the same language, in versions sometimes labelled as "easy-to-read", or "modern" language etc. In English, for example, it is clear that St. James' version is the most difficult to read ($G = 57.2$) but it loads less reader's short-term memory ($I_P = 5.91$) than the Contemporary English Edition (CEV) ($G = 66.8; I_P = 8.28$). In German, the versions tend be much closer, even Luther's, so that they seem to address very similar audiences.

The spread of the values within the same language can be a sizeable fraction of the overall range calculable from Table 1 and Table 2. For example, for English, the spread in W 8% is to be compared to the overall (Table 1) 61.9%; for S, the spread 75.3% is to be compared to 106.9%. Therefore, an English translation can be confused, mathematically, with the translation in another language.

[39] https://classic.biblegateway.com/versions/

Table 5. Matthew's Gospel (28 chapters). Total number of words W and sentences S; average number of words per sentence P_F, average number of words per interpunction I_P, average number of interpunctions per sentence M_F and readability index G for the indicated translations. The source of the unnoted translations is reported in **Table 1**. The range (%) is defined as the ratio between the difference between maximum and minimum values (range) and the Greek value, multiplied by 100.

Version	W	S	P_F	I_P	M_F	G
Greek	18,121	914	20.66	7.23	2.86	59.1
English	22,000	1247	17.88	6.91	2.61	61.8
CEV[40]	23,444	1728	13.64	8.28	1.67	67.8
St James[41]	23,397	1040	23.51	5.91	3.98	57.2
Range (%)	8.0	75.3	47.8	32.8	80.8	17.9
German	21,424	1324	16.69	5.80	2.90	61.4
NGU-DE[42]	23,122	1534	15.28	5.85	2.62	62.0
Luther[43]	21,998	1211	18.71	5.93	3.17	60.2
Range (%)	9.4	23.0	16.6	1.8	19.2	3.0
Polish	17,650	1563	11.61	4.54	2.56	59.3
Slowo[44]	17,211	1677	10.55	4.83	2.19	61.0
UBG[45]	17,651	1299	13.89	4.58	3.05	55.7
Range (%)	2.4	41.4	16.2	4.0	30.1	9.0
Russian	16,786	956	18.33	4.12	4.46	58.9
CARS[46]	18,243	1359	13.65	4.50	3.04	63.6
ERV-RU[47]	18,395	1353	13.90	4.65	3.00	63.5
Range (%)	8.9	44.1	22.7	7.3	51.0	8.0
Spanish	21,217	1232	18.22	6.12	2.99	60.5
CST[48]	21,318	1392	16.07	6.41	2.52	62.6
TLA[49]	25,367	1630	15.47	7.00	2.22	62.5
Range (%)	22.9	43.5	13.3	12.2	26.9	3.6
Swedish	21,552	1445	15.10	7.50	2.02	63.1
SFB15[50]	20,676	1409	15.10	7.56	2.00	64.3
SV1917[51]	22,503	1182	19.72	6.59	3.00	57.9
Range (%)	5.2	28.8	22.4	13.4	35.0	10.8

[40]https://classic.biblegateway.com/versions/Contemporary-English-Version-CEV-Bible/#booklist
[41]https://classic.biblegateway.com/versions/New-King-James-Version-NKJV-Bible/#booklist
[42]https://classic.biblegateway.com/versions/Neue-Genfer-%C3%9Cbersetzung-NGU/#booklist
[43]https://classic.biblegateway.com/versions/Luther-Bibel-1545-LUTH1545/#booklist
[44]https://classic.biblegateway.com/versions/S%C5%82owo-%C5%BBycia-SZ/#booklist
[45]https://classic.biblegateway.com/versions/Updated-Gda%C5%84sk-Bible-UBG/#booklist
[46]https://classic.biblegateway.com/versions/-CARS/#booklist
[47]https://classic.biblegateway.com/versions/Russian-Bible-Easy-to-Read-Version-ERV-RU/#booklist
[48]https://classic.biblegateway.com/versions/Nueva-Version-Internacional-Castilian-Biblia-CST/#booklist
[49]https://classic.biblegateway.com/versions/Traducci%C3%B3n-en-lenguaje-actual-TLA-Biblia/#booklist
[50]https://classic.biblegateway.com/versions/Svenska-Folkbibeln-2015-SFB15-Bible/#booklist
[51]https://classic.biblegateway.com/versions/Svenska-1917-SV1917/#booklist

In conclusion, it is clear that each NT different translation within the same language addresses different audiences, as it can be noticed from the range of the linguistic parameters, but, more interestingly, a translation in a language can be confused, mathematically, with the translation in another language. In other words, this preliminary sampling seems to confirm that language does not play the only role in translation, but that this role has to be shared mainly with reader's reading ability (*i.e.*, P_P, G) and short-term memory (I_P).

10. Literary Text Translations: *Treasure Island*

Another question arises: Are the above results only applicable to NT translations, or can they be also applied to translations of literary texts, such as novels? In this section we show, preliminarily with just one example, that novels tend to show similar statistics, but with more constraints on the translations than those found in the NT translations.

We have done the following exercise. We have studied the translations of *Treasure Island* (by R.L. Stevenson) from the original English text to Italian, French and German, by considering each chapter as text unit (34 chapters).

The comparison to the NT translation must be done, of course, by starting first with the English version of the NT and then studying its translations. Only after this study, we can consider *Treasure Island* as input text and calculate the same statistics. Therefore, we take the English NT as the reference (input) language and Italian, French and German as output languages, as if these NT versions were obtained by translating the English text, not the original Greek text. This hypothesis assumes, of course, that if the Italian, French and German translators had started from the English version of the NT, they would have ended up with the same text translated from Greek. This might be reasonable, although not directly controllable. We show below that the assumption can be justified. Table 6 reports the statistics concerning *Treasure Island* original text and its translations.

Table 7, Table 8 report the results on channel capacity obtained by considering English as the original NT text, while Table 9, Table 10 report the results on channel capacity concerning the direct translations of *Treasure Island* to Italian, French and German. We notice that the Italian translation uses the least number of words and sentences, and has also the highest correlation coefficients for all variables; therefore, its channels have also the largest capacities. In other words, the Italian translation is, mathematically, the closest to the English text, which appears surprising if we consider the different linguistic family.

Let us examine the single channels. In the words channel n_W we notice that the slope m and correlation coefficient r of the three languages are about the same in both cases (Table 7 and Table 9), therefore our hypothesis, mentioned in the previous paragraph, on the translation of the English NT to the other languages is justified. More interesting, the channel capacity is about the same in both cases and very close to the maxima given by Equation (15).

Table 6. (a) *Treasure Island* statistics. Key: total; average (standard deviation); slope m and correlation coefficient r between the translation and the original English text. (b) *Treasure Island* statistics. Average (standard deviation); slope m and correlation coefficient r between the translation and the original English text.

(a)

Language	Words	m	r	Sentences	m	r
English	68,033; 2001.0 (302.3)	1	1	3824; 112.5 (31.4)	1	1
Italian	64,603; 1900.1 (294.6)	0.950	0.985	3805; 111.9 (30.2)	1.181	0.904
French	68,818; 2024.1 (334.6)	1.013	0.982	4054; 119.2 (30.8)	1.253	0.874
German	72,119; 2121.1 (332.6)	1.060	0.970	4111; 120.9 (31.5)	0.889	0.833

(b)

Language	P_F	m	r	I_P	m	r	M_F	m	r
English	18.9 (9.8)	1	1	6.05 (1.86)	1	1	3.09 (0.77)	1	1
Italian	17.9 (8.4)	0.95	0.907	6.52 (1.68)	1.024	0.900	2.72 (0.73)	0.796	0.725
French	17.9 (8.4)	1.013	0.882	6.11 (1.62)	0.959	0.927	2.88 (0.66)	0.842	0.665
German	18.3 (7.6)	1.060	0.643	5.96 (1.53)	1.025	0.861	3.05 (0.75)	1.107	0.522

Table 7. NT statistics on channel capacity (bits per symbol): Translations from English to Italian, French and German; n_W and n_S channels; n.a. stands for "not applicable".

Language	Words	n_W channel				Sentences	n_S channel			
		m	r	C	C/C_{max}		m	r	C	C/C_{max}
English	122,641	1	1	n.a.	n.a.	6590	1	1	n.a.	n.a.
Italian	112,943	0.918	0.993	2.852	0.923	6396	0.963	0.951	1.728	0.982
French	133,050	1.077	0.984	2.270	0.904	7258	1.076	0.945	1.497	0.888
German	117,269	0.952	0.979	2.330	1.000	7069	1.064	0.952	1.607	0.907

Table 8. NT statistics on channel capacity (bits per symbol): Translations from English to Italian, French and German; P_F and I_P channels.

Language	P_F channel				I_P channel			
	m	r	C	C/C_{max}	m	r	C	C/C_{max}
Italian	0.757	0.602	0.478	0.703	0.846	0.536	0.318	0.503
French	0.768	0.619	0.500	0.719	0.999	0.676	0.441	0.585
German	0.713	0.721	0.736	0.906	0.779	0.674	0.597	0.795

Table 9. *Treasure Island* statistics on channel capacity (bits per symbol): Translations from English to Italian, French and German; n_W and n_S channels.

Language	n_W channel				n_S channel			
	m	r	C	C/C_{max}	m	r	C	C/C_{max}
Italian	0.950	0.985	2.542	0.995	1.181	0.904	0.982	0.729
French	1.013	0.982	2.375	0.978	1.253	0.874	0.749	0.627
German	1.060	0.970	1.930	0.927	0.889	0.833	0.959	0.916

Table 10. *Treasure Island* statistics on channel capacity (bits per symbol): Translations from English to Italian, French and German; P_F and I_P channels.

Language	P_F channel				I_P channel			
	m	r	C	C/C_{max}	m	r	C	C/C_{max}
Italian	0.950	0.907	1.301	0.953	1.024	0.899	1.165	0.884
French	1.013	0.882	1.070	0.870	0.959	0.927	1.460	0.967
German	1.060	0.642	0.349	0.487	1.025	0.861	0.946	0.829

In the sentences channel n_S, on the contrary, m and r of the three languages are significantly different in the two cases. This is, of course, confirmed by the different capacities. This trend is further enhanced in the P_F and I_P channels (Table 8 and Table 10), another evidence that, as we pass from words to sentences, to P_F and to I_P (or M_F), each translation has quite different ways of using interpunctions for their intended readers, therefore matching more reader's reading ability and short-term memory capacity.

Finally, it is very interesting to notice in n_W and n_S channels (Table 7 and Table 9), that the NT translation, mathematically, is more accurate and respectful of the original Greek text than the translation of *Treasure Island*. On the contrary, in P_F and I_P channels, *Treasure Island* translations are more accurate than NT translations because, very likely, all dialogues must be strictly respected in any translation.

In conclusion, the statistics of words and sentences of a novel seems to be similar to those found in the NT translations. For example, the ranking of the number of sentences, from minimum to maximum, is the same both in the NT and in the *Treasure Island* translations: Italian, English, French, German. It is almost the same for words, namely, Italian, German, English, French for the NT translations; Italian, English, German, French for *Treasure Island* translations. The translation of a novel seems to be more respectful of the original text than the NT translations for what concerns P_F and I_P, mainly because the translators must consider the presence of dialogues, whose fraction of the total text can be, however, largely variable within novels, according to author's style etc. Because these results refer to just one particular case, they should be further assessed with other literary (novels) translations, a study well beyond the aim of this paper.

11. A General Theory of Translation: From Any Language to Any Other Language

It is possible to extend the statistical theory outlined in the previous sections in such a way to arrive at a general theory of translation applicable to any alphabetical language. By knowing the statistics of the various linguistic variables studied in the previous sections—obtained in the translation channel from Greek to other languages—it is possible as we show below, to estimate the statistics obtainable in the translation channel from any language to any other language of those listed in Table 1. The necessary data for extending the theory are those

reported in Table 3, Table 4.

The theory can also be applied to channels of texts belonging to the same language (not showing for brevity): for example, the channel that transforms words into sentences in a text can be compared to the channel that transforms words into sentences in a different text, both written in the same language. This comparison can be useful to study how texts of the same author may have changed over time, or to compare texts of different authors.

Figure 20 shows, schematically, the block diagram of the direct channels from language Y_k ($k = 1:36$, Greek in Figure 20) to language Y_j (channel $Y_k \to Y_j$; $j = 1:36$) and the flow chart of the reverse channels, from any language Y_j to the same language Y_k (channels $Y_j \to Y_k, k = 1:36$, Greek in Figure 20). In other words, in the direct channel the translation is from a single language (Greek, or Latin, or Esperanto etc.) to another language, therefore, if the starting language is Greek, the translations are those discussed in the previous sections. In the reverse channel the output language is the same for all translations, therefore if the output language is Greek, the translations are from input languages Latin, Esperanto etc. So far, we have studied only one possible direct channel (from Greek to the other languages) and none of the reverse channels. In this section we study all possible direct and reverse channels for proposing a statistical general theory of translation.

We first calculate the noise-to-signal power ratio obtainable in the general theory from the data reported in Table 3, Table 4. After, we show that direct and reverse channels concerning any couple of languages are not symmetric.

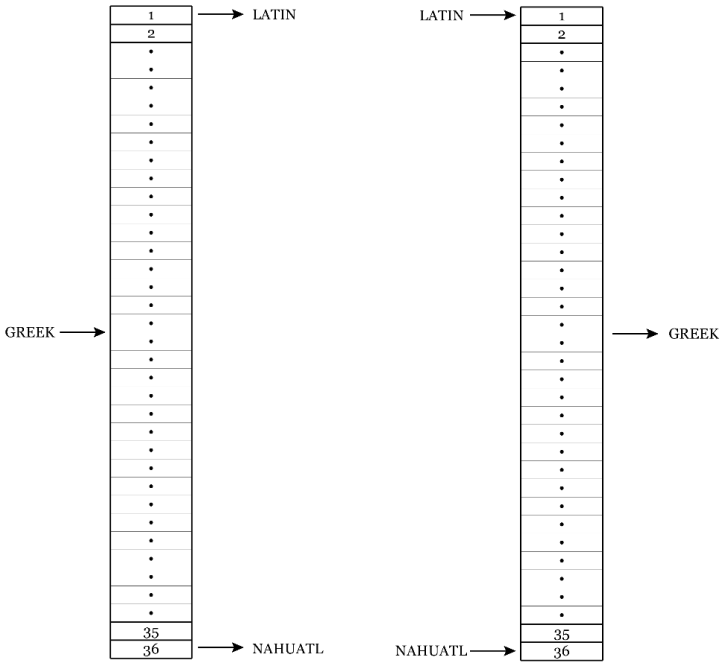


Figure 20. **Left**: Direct Channel: translation from a language (common input) to all other languages (output). **Right**: Reverse channel: translation from all languages (different input) to one language (common output).

11.1. Noise-to-Signal Power Ratio

Let us consider two languages Y_k and Y_j, and let us refer to Greek explicitly as language X. With reference to the ideal channel whose output is X (self-translation), we have found that the same variable of languages k and j are related by linear relationships with the corresponding Greek variable x:

$$y_k = m_k x + n_k \tag{19a}$$

$$y_j = m_j x + n_j \tag{19b}$$

In Equation (19) n_k and n_j are the noise sources added to the regression lines $y = mx$. The slope m is the source of the regression noise—because $|m| \neq 1$—the correlation coefficient r is the source of the correlation noise—because $|r| < 1$—as discussed in Section 4. For example, in the words channel between Greek and English, $m = 1.225$ and $r = 0.986$ (Table 3).

Let us refer to the 36 possible translations from language $k = 1:37$—including Greek—to language j. In other words, language k plays now the role played before by Greek. By eliminating x, *i.e.* Greek, from Equation (19), we get the linear relationship between the input language k and the output language j:

$$y_j = \frac{m_j}{m_k} y_k - \frac{m_j}{m_k} n_k + n_j \tag{20}$$

Compared to the reference language y_k, the slope is given by:

$$m_{kj} = \frac{m_j}{m_k} \tag{21}$$

Therefore, the regression noise-to-signal power ratio, R_m, of the channel is readily found, according to Equation (3), as:

$$R_m = \left(m_{kj} - 1\right)^2 \tag{22}$$

Notice that R_m depends only the known slopes of the translations from Greek (Table 3).

Let us calculate the correlation noise-to-signal power ratio, R_r. To apply Equation (6), we must insert the unknown correlation coefficient $r_{kj} = r_{jk} = r$ between y_j and y_k due, of course, to the two noise sources in Equation (20). We can calculate its value from the correlation coefficients r_k and r_j reported in Table 3. First, we notice that the total noise added to the regression line relating the output variable y_j to the input variable y_k is given by:

$$n_{j,tot} = -m_{kj} n_k + n_j \tag{23}$$

As we can see from (22), the two noise sources are correlated, with unknown correlation coefficient r. Let s_{nk}^2 and s_{nj}^2 be the single noise powers, then the total noise power $s_{n,j,tot}^2$ due to $n_{j,tot}$ is given by ([39], p.127):

$$s_{n,j,tot}^2 = m_{kj}^2 s_{nk}^2 + s_{nj}^2 + 2 r m_{kj} s_{nk} s_{nj} \tag{24}$$

Equation (24) has a geometric representation [39]. It can be seen as an appli-

cation of the law of cosine to the vectors $m_{kj}s_{nk}$ and s_{nj}, which form the angle $\theta_{kj} = \arccos(r_{kj})$ between them. By applying this representation also to the vectors s_{nk} and s_x (Greek) forming the angle $\theta_k = \arccos(r_k)$ and to the vectors s_{nj} and s_x, forming the angle $\theta_j = \arccos(r_j)$, the angle θ_{kj} is given by $\theta_{kj} = |\theta_j - \theta_k|$, therefore r is given by:

$$r = \cos\left|\arccos(r_j) - \arccos(r_k)\right| \tag{25}$$

Now, by Equation (6), the correlation noise-to-signal power ratio in the translation channel from language k to language j is given by:

$$R_r = \frac{1-r^2}{r^2} m_{kj}^2 \tag{26}$$

In conclusion, the total noise-to-signal power ratio in the translation channel from language k to language j, for a given stochastic variable, is given by:

$$R = (m_{kj} - 1)^2 + \frac{1-r^2}{r^2} m_{kj}^2 \tag{27}$$

Figures 21-23 show the geometrical representation of R_m and R_r in the first Cartesian quadrant as discussed in Section 4, for all linguistic variables. Notice that the regression lines from Greek to other languages, drawn from Figure 7 and Figure 8, are approximately upper bounds to the general theory in the words, sentences and interpunctions channels. Moreover, also for the other variables, Greek direct and reverse channels are noisier than other languages. In other words, modern languages and Latin are statistically closer to each other than to Greek. We also notice two different features: the words n_W, sentences n_S and interpunctions n_I channels are mostly dominated by R_m, because for most languages $X > Y$, *i.e.* $R_m > R_r$. This result underlines, again, the greater freedom used in these channels in sizing the number of words, sentences and interpunctions, whose average values may vary substantially (Table 1 and Table 2), while keeping very high correlation coefficients (Table 3). In the words channel for example $0.881 \le m \le 1.518$, and $0.949 \le r \le 0.994$. On the contrary, in the channels concerning the deep-language variables P_F, I_P (with some exceptions), M_F, and the readability index G, we mostly observe $X < Y$, *i.e.* $R_m < R_r$. In the P_F channel, for example, in Table 3 we read $0.529 \le m \le 1.0$ and $0.363 \le r \le 0.883$, with a significant impact on the noise-to-signal power ratio.

From Figures 21-23 we can calculate direct and reverse channels capacities. Figure 24 shows the scatterplots between C_{kj} (direct channel) and C_{jk} (reverse channel) for some languages in the words channel n_W.

Figure 25 shows the scatterplot of the averages of all languages for the words channel. Notice that the perfect even symmetry around the 45˚ line is due to how the table from which the data are taken is built. However, the interesting point is the very small data scattering around the 45˚ line, which yields a small $\langle \Delta C_{kj} \rangle = \langle C_{jk} - C_{kj} \rangle = \langle C_{jk} \rangle - \langle C_{kj} \rangle$. Similar scatterplots are also obtained for the

other channels (see Appendix F).

These scatterplots show that direct and reverse channels are not very different. Although $C_{kj} \neq C_{jk}$, as we establish in the next subsection, they are, however, very similar for all variables and languages, regardless of their absolute value. In other words, a common underlying structure emerges from considering channel capacities, which seems to govern textual/verbal communication channels defined here, as we can see in Figure 25. In Appendix F we show results for the other linguistic channels.

In the next subsection we show that $C_{kj} \neq C_{jk}$.

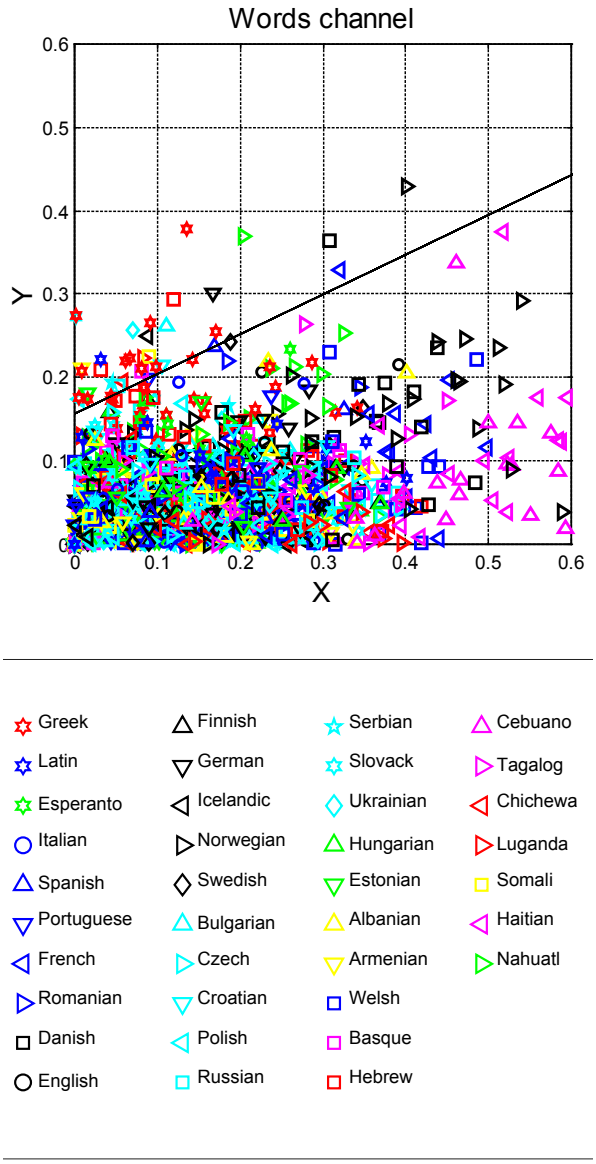


Figure 21. Upper: Scatterplot between $X = \sqrt{R_m}$ and $Y = \sqrt{R_r}$ in n_w channels. The origin represents the ideal channel. For each language 36 identical symbols are shown, because it is the common output of the translations from the remaining 36 languages. The regression line is redrawn from Figure 7. **Lower:** symbols caption.

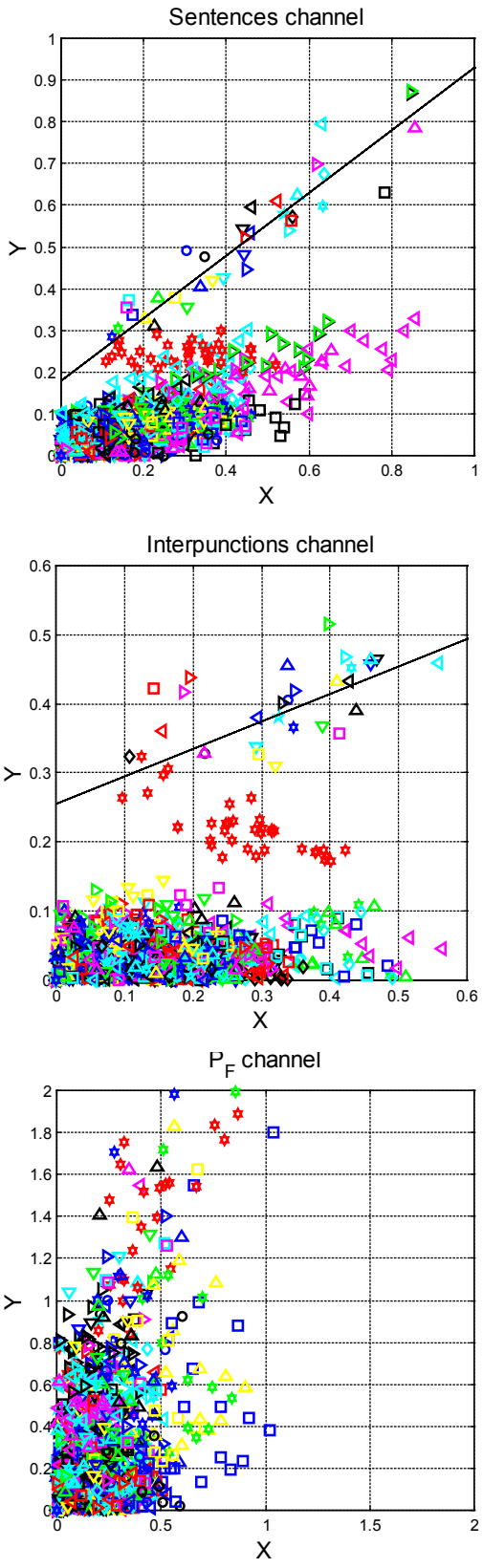


Figure 22. Scatterplot between $X=\sqrt{R_m}$ and $Y=\sqrt{R_r}$. **Upper:** n_S channels. **Middle:** n_I channels. **Lower:** P_F channels.

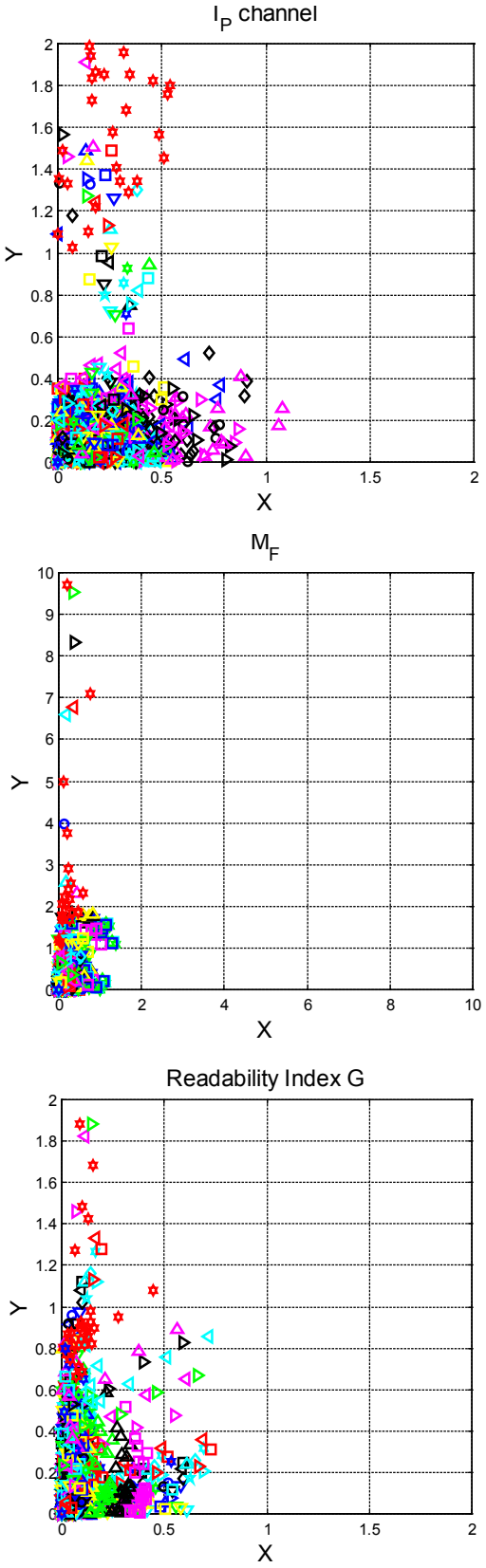


Figure 23. Scatterplot between $X=\sqrt{R_m}$ and $Y=\sqrt{R_r}$. **Upper:** I_P channels. **Middle:** M_F channels. **Lower:** G channels.

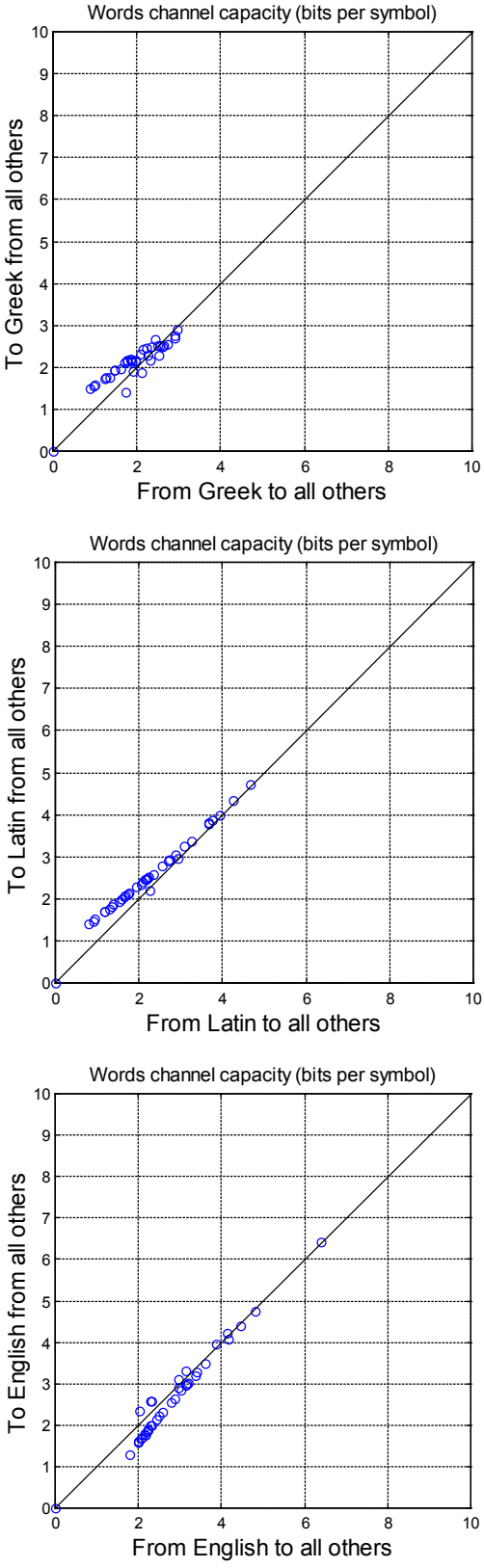


Figure 24. n_W channels. **Upper:** Scatterplot between direct channel capacity (from … to) and the reverse channel capacity (to … from) for Greek. **Middle:** (to … from) for Latin. **Lower:** (to … from) for English.

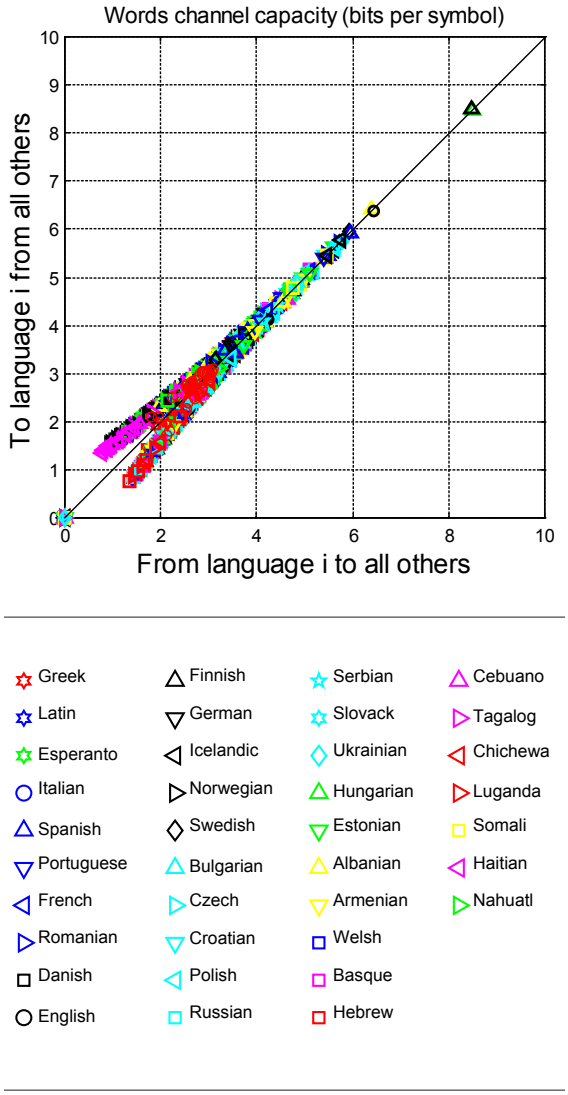


Figure 25. Upper: Scatterplot between direct channel capacity (from … to) and the reverse channel capacity (to … from) for all languages, n_W channel. The origin represents the ideal channel. The large red symbol is the overall average value. **Lower:** symbols caption.

11.2. Direct and Reverse Channels Are Not Symmetric

Are direct and reverse channels concerning a couple of languages, e.g. translations from Greek to English and from English to Greek, symmetric? We can answer to this question by considering the channel capacity.

The specific question becomes now: Is the capacity C_{kj} (bits per symbol) of the (direct) channel from language k to language j, equal to the capacity C_{jk} of the (reverse) channel from language j to language k? In other words, can the two languages be exchanged in the input-output relationship without changing the statistical characteristics of the translation channel? According to communication theory [26], this happens in telecommunication channels affected by additive white Gaussian noise, but this is not true in translation channels, as we show

next.

We establish now that any couple of direct and reverse channels are *not* symmetric, unless $m_k = m_j$ and $r_k = r_j$, a case never found. The reason for this asymmetry is because the noise added to any ideal (self-translation) channel to get the text in another language is statistically always different.

According to Equations (12) and (27), and recalling that $r_{ij} = r_{ji} = r$, the two channel capacities are equal if:

$$\left(\frac{m_j}{m_k}-1\right)^2+\frac{1-r^2}{r^2}\left(\frac{m_j}{m_k}\right)^2=\left(\frac{m_k}{m_j}-1\right)^2+\frac{1-r^2}{r^2}\left(\frac{m_k}{m_j}\right)^2 \tag{28}$$

Let $x = m_j/m_k$. After standard algebraic passages, we get following solution for the unknown correlation coefficient:

$$r=\mp\sqrt{\frac{x^2+1}{2x}} \tag{29}$$

To yield real values, the radicand in Equation (29) must be positive, and to yield a correlation coefficient must be less than 1, therefore we get the range:

$$0\le\frac{x^2+1}{2x}\le 1 \tag{30}$$

The lower limit in (31) is always satisfied because $x^2+1>0$; the upper limit gives:

$$x^2-2x+1=(x-1)^2\le 0 \tag{31}$$

The inequality (31) is never satisfied, unless $x = 1$, therefore only if $m_j = m_k$, in which case, from Equation (29) $r = \pm 1$. In other words, in translation channels $C_{jk} \neq C_{kj}$. Only in the ideal channel (self-translation) $C_{jk} = C_{kj} = \infty$. In the next subsection we assess how large the capacity difference is, in other words, how asymmetric direct and reverse channels are.

11.3. Direct and Reverse Channels Capacity Difference

Figures 26-28 show ΔC_{kj} main statistics, for all couples of direct and reverse channels—and for the same linguistic variable—for each language, by drawing, as a function of $\langle \Delta C_{kj} \rangle$, the standard deviation $\sigma_{\Delta C}$, the root mean square (RMS) value (bits per symbol) and its relative (normalized) value RMS (%)—the latter obtained by dividing RMS of ΔC_{kj} by the average direct channel capacity $\langle C_{kj} \rangle$. Table 11 reports averages.

Several interesting observations can be done. First, we notice that $\langle \Delta C_{kj} \rangle$, $\sigma_{\Delta C}$ and RMS vary in about the same range. The average value, for example is approximately always in the range $-0.4 \lesssim \langle \Delta C_{kj} \rangle \lesssim +0.4$ (bits per symbol), regardless of the variable. Only Greek is clearly distinct from the other languages, with larger values. The standard deviation is even more stable as $\sigma_{\Delta C} \approx 0.2$ (bits per symbol) in most cases. Only RMS has larger variations, between 0.2 and 0.6 (bits per symbol). As already noticed, Latin and modern languages are closer to each other than to Greek.

On the contrary, the variations of the normalized RMS (%) are significantly different. In the words channel RMS varies between 10% and 30%, and similarly for the sentences channel (10% to 40%) and interpunctions channel (10% to 20%); on the contrary RMS varies in a larger range in the deep-language channels P_F, I_P and M_F, up to 300%.

We can rank the channels according to the normalized RMS (%). Table 12 shows its overall average. The least variable channel is the readability channel, followed by the interpunction channel, the words and sentences channels, then the deep-language channels, therefore confirming that these latter variables are treated by translators with fewer constraints than the number of words or sentences, unless dialogues have to be respected, as seen with *Treasure Island* translations. In other words, in the NT translations differences are mainly due to specific linguistic variables, not to the particular language.

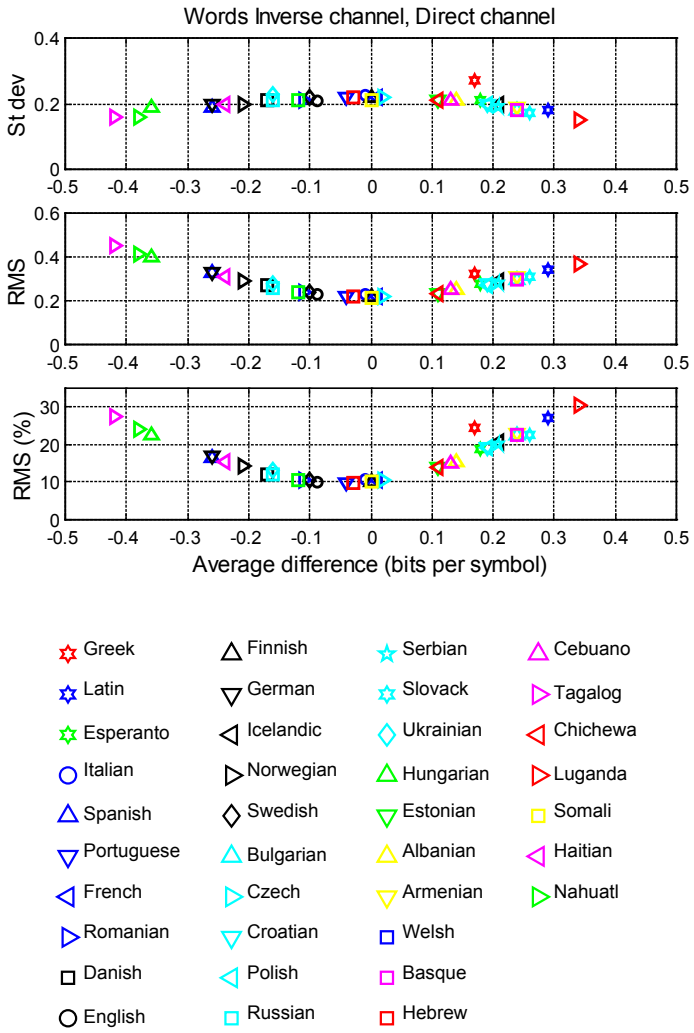


Figure 26. Standard deviation, RMS and normalized RMS (%) values versus average capacity difference of the reverse and direct n_W channels.

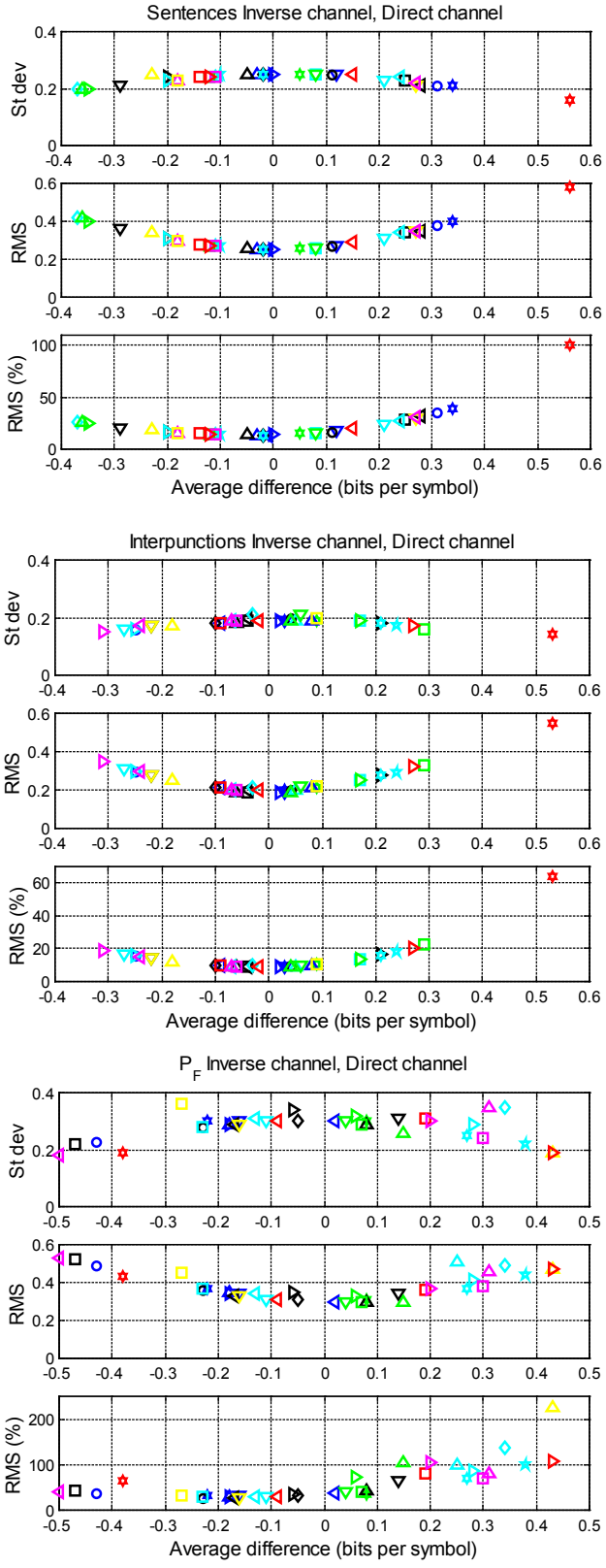


Figure 27. Standard deviation, RMS and normalized RMS (%) values versus the average capacity difference of the reverse and direct channels. **Upper:** n_S channels. **Middle:** n_I channels. **Lower:** P_F channels.

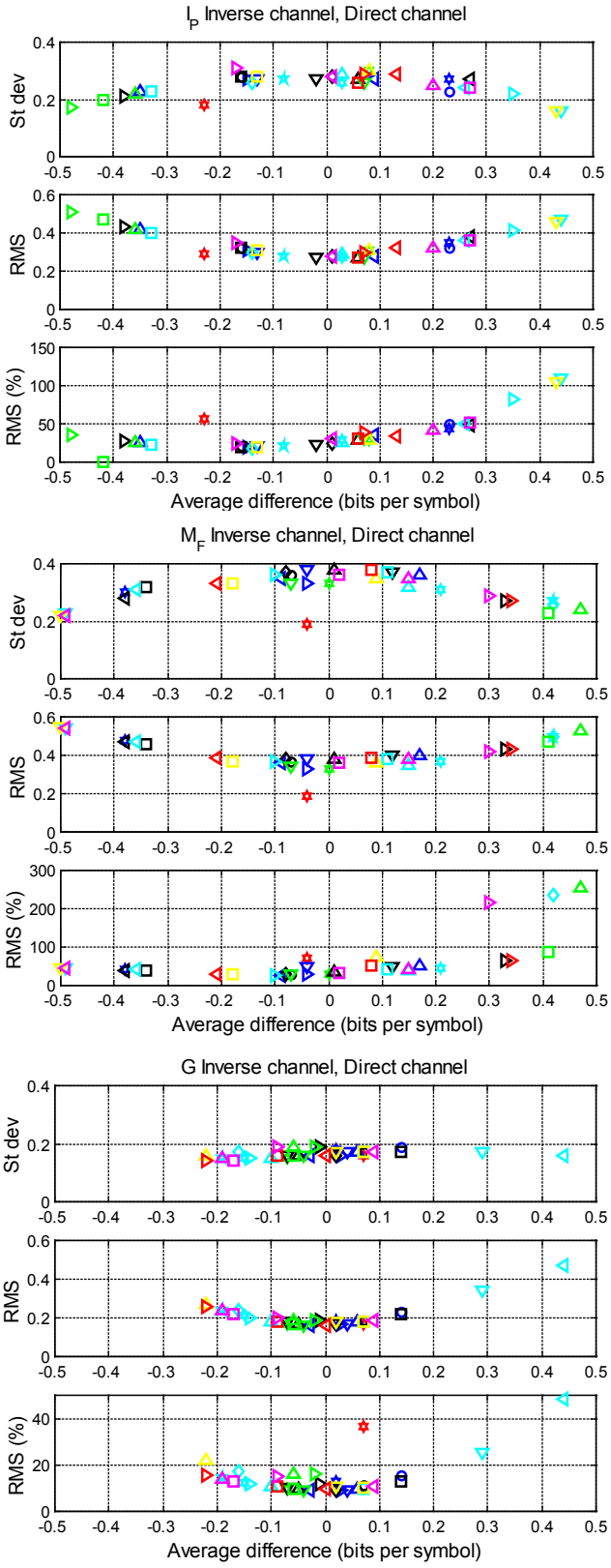


Figure 28. Standard deviation, RMS and normalized RMS (%) values versus the average capacity difference of the reverse and direct channels. **Upper:** I_P channels. **Middle:** M_F channels. **Lower:** G channels.

Table 11. Direct and reverse channels average statistics.

Channel	$C_{ave} = \langle\langle (C_{kj} + C_{jk}) \rangle\rangle / 2$	s_C	$100 \times s_C / C_{ave}$
Words per chapter n_P	2.67	0.27	10.1
Sentences per chapter n_S	2.62	0.33	12.6
Interpunctions per chapter n_I	3.02	0.26	8.6
Words per sentence P_F	1.79	0.39	21.8
Words per interpunctions (word interval) I_P	2.09	0.35	16.7
Interpunctions (word intervals) per sentence M_F	1.59	0.43	27.0
Readability index per chapter G	2.42	0.23	9.5

Table 12. Channels ranking according to the overall normalized RMS (%).

	Interpunctions	G	Words	Sentences	I_P	P_F	M_F
RMS (%)	13.7	15.9	16.7	22.4	36.0	59.9	73.2

12. Conclusions

We have proposed a unifying statistical theory of translation, based on communication theory, which involves linguistic stochastic variables, some of which are not considered by scholars. Its main mathematical characteristics have emerged by studying the translation of most NT books.

When a text written in a language is translated into another language, all linguistic variables do numerically change. To study these apparently chaotic data we have characterized any translation as a complex communication channel affected by "noise", studied according to Communication Theory applied for the first time to this channel. The new theory deals with aspects of languages more complex than those currently considered in machine translations. The input language is the "signal", the output language is a "replica" of the input language, but largely perturbed by noise. For the output language, this noise is indispensable for conveying the meaning of the input language to its readers. To study these channels, we have defined a suitable noise-to-signal power ratio and applied a geometrical representation.

All channels studied are differently affected by translation noise. The more accurate channel is the word channel n_W, a finding that seems reasonable. It emerges that humans seem to express a given meaning with a number of words—*i.e.* finite strings of abstract signs (characters)—which cannot vary so much even if some languages do not share a common ancestor. On the contrary, the number of sentences and especially their length in words, *i.e.* P_F, are treated more freely by translators. P_F, affects readability indices very much, therefore this variable tends to be better matched to the intended readers, with specific reading ability.

Independently of the different parallel channels (one for each variable), the correlation noise (due to a regression line slope $m \neq 1$) is mostly larger than the regression noise (due to a regression correlation coefficient $r < 1$), therefore in-

dicating that every translation tries as much as possible to be not biased, but it cannot avoid being decorrelated, with correlation coefficients which approximately decrease from words, to sentences, to interpunctions and down to the deep-language variables P_F, I_P, M_F and C_P.

Different translations of the NT within the same language, mathematically, can be quite different and they can even seem to belong to different languages. In other words, in language translations differences are mainly due to specific linguistic variables, not to the particular language. Clearly, they are matched to different audiences, an aspect not explicitly considered in machine translations.

Besides the noise-to-signal power ratio, communication channels can be also characterized by the channel capacity (bits per symbol, the latter suitably defined). This parameter can be relatively large, very close to the maximum value obtainable, for n_W, n_S and n_I channels, less for P_F, I_P, M_F channels. We have found that the NT translations are similar to translations of literary texts, as shown for the novel *Treasure Island* translated from English to Italian, French and German for n_W, n_S and n_I channels. On the contrary, the translation of novels seems to set more stringent constraints on the translators for P_F, and I_P, channels because dialogues must be strictly maintained. A topic to be further researched.

The number of words per interpunctions I_p varies in the same range of the short-term memory capacity. Drawn against the number of words per sentence P_F, I_p tends to saturate to a horizontal asymptote as P_F increases because, even though sentences get longer, I_p cannot get larger than about the upper limit of Millers' law, because of the constraints imposed by readers' short-term memory capacity.

We have defined a formula for the readability index of any alphabetical languages, based on a calque of the readability formula used in Italian, both for providing it to languages that have none, and also for estimating, on common grounds, the readability of texts belonging to different languages/translations.

Finally, we have extended the statistical theory outlined before to a general theory of translation applicable to any alphabetical language, even to texts written in the same language. The general theory shows that direct and reverse channels are not symmetric.

In conclusion, a common underlying statistical structure, governing human textual/verbal communication channel—not defeated by the mythical biblical Tower of Babel—seems to emerge from the findings. The main result is that the statistical and communication characteristics of a text, and its translations into other languages, seem to depend not only on the particular language—mainly through the number of words and sentences—but also on the particular translation because the text is very much characterized by the reading abilities and short-term memory capacity of the intended readers, aspects not explicitly considered in machine translations. These conclusions seem to be everlasting because applicable also to ancient Roman and Greek readers. A future research should extend the general theory to non-alphabetical languages.

Acknowledgements

The author wishes to warmly thank all those scholars who, with continuous great care and dedication, keep online the texts of the Bible in many languages, for the benefit of everyone.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Catford, J.C. (1965) A Linguistic Theory of Translation. An Essay in Applied Linguistics. Oxford University Press, Oxford.

[2] Munday, J. (2008) Introducing Translation Studies: Theories and Applications. 2nd Edition, Routledge, London. https://doi.org/10.4324/9780203869734

[3] Proshina, Z. (2008) Theory of Translation. 3rd Edition, Far Eastern University Press, Vladivostok.

[4] Trosberg, A. (2000) Discourse Analysis as Part of Translator Training. *Current Issues in Language and Society*, **7**, 185-228. https://doi.org/10.1080/13520520009615581

[5] Tymoczko, M. (1999) Translation in a Post-Colonial Context: Early Irish Literature in English Translation. St Jerome, Manchester.

[6] Warren, R. (1989) The Art of Translation: Voices from the Field. North-Eastern University Press, Boston.

[7] Williams, I. (2007) A Corpus-Based Study of the Verb Observar in English-Spanish Translations of Biomedical Research Articles. *Target*, **19**, 85-103. https://doi.org/10.1075/target.19.1.06wil

[8] Wilss, W. (1996) Knowledge and Skills in Translator Behaviour. John Benjamins, Amsterdam and Philadelphia. https://doi.org/10.1075/btl.15

[9] Wolf, M. and Fukari, A. (2007) Constructing a Sociology of Translation. John Benjamins, Amsterdam and Philadelphia. https://doi.org/10.1075/btl.74

[10] Gamallo, P., Pichel, J.R. and Alegria, I. (2020) Measuring Language Distance of Isolated European Languages. *Information*, **11**, 181. https://doi.org/10.3390/info11040181

[11] Barbançon, F., Evans, S., Nakhleh, L., Ringe, D. and Warnow, T. (2013) An Experimental Study Comparing Linguistic Phylogenetic Reconstruction Methods. *Diachronica*, **30**, 143-170. https://doi.org/10.1075/dia.30.2.01bar

[12] Collins-Thompson, K. (2014) Computational Assessment of Text Readability: A Survey of Past, in Present and Future Research, Recent Advances in Automatic Readability Assessment and Text Simplification, ITL. *International Journal of Applied Linguistics*, **165**, 97-135. https://doi.org/10.1075/itl.165.2.01col

[13] Bakker, D., Muller, A., Velupillai, V., Wichmann, S., Brown, C.H., Brown, P., Egorov, D., Mailhammer, R., Grant, A. and Holman, E.W. (2009) Adding Typology to Lexicostatistics: A Combined Approach to Language Classification. *Linguistic Typology*, **13**, 169-181. https://doi.org/10.1515/LITY.2009.009

[14] Grzybeck, P. (2007) History and Methodology of Word Length Studies, Contributions to the Science of Text and Language. Spinger, Dordrecht, 15-90.

https://doi.org/10.1007/978-1-4020-4068-9_2

[15] Petroni, F. and Serva, M. (2010) Measures of Lexical Distance between Languages. *Physica A*: *Statistical Mechanics and Its Applications*, **389**, 2280-2283. https://doi.org/10.1016/j.physa.2010.02.004

[16] Gómez-Adorno, E, Sidorov, G., Pinto, D. Vilariño D. and Gelbukh, A. (2016) Automatic Authorship Detection Using Textual Patterns Extracted from Integrated Syntactic Graphs. *Sensors*, **16**, 19 p. https://doi.org/10.3390/s16091374

[17] Carling, G., Larsson, F., Cathcart, C., Johansson, N., Holmer, A., Round, E. and Verhoeven, R. (2018) Diachronic Atlas of Comparative Linguistics (DiACL)—A Database for Ancient Language Typology. *PLoS ONE*, **13**, e0205313. https://doi.org/10.1371/journal.pone.0205313

[18] Gao, Y., Liang, W., Shi, Y. and Huang, Q. (2014) Comparison of Directed and Weighted Co-Occurrence Networks of Six Languages. *Physica A*: *Statistical Mechanics and Its Applications*, **393**, 579-589. https://doi.org/10.1016/j.physa.2013.08.075

[19] Liu, H. and Cong, J. (2013) Language Clustering with Word Co-Occurrence Networks Based on Parallel Texts. *Chinese Science Bulletin*, **58**, 1139-1144. https://doi.org/10.1007/s11434-013-5711-8

[20] Gamallo, P., Pichel, J.R. and Alegria, I. (2017) From Language Identification to Language Distance. *Physics A*, **484**, 162-172. https://doi.org/10.1016/j.physa.2017.05.011

[21] Pichel, J.R., Gamallo, P. and Alegria, I. (2019) Measuring Diachronic Language Distance Using Perplexity: Application to English, Portuguese, and Spanish. *Natural Language Engineering*, **26**, 433-454. https://doi.org/10.1017/S1351324919000378

[22] Eder, M. (2015) Visualization in Stylometry: Cluster Analysis Using Networks. *Digital Scholarship in the Humanities*, **32**, 50-64. https://doi.org/10.1093/llc/fqv061

[23] Brown, P.F., Cocke, J., Della Pietra, A., Della Pietra, V.J., Jelinek, F., Lafferty, J.D., Mercer, R.L. and Roossin, P.S. (1990) A Statistical Approach to Machine Translation. *Computational Linguistic*, **16**, 79-85.

[24] Koehn, F., Och, F.J. and Marcu, D. (2003) Statistical Phrase-Based Translation. *Proceedings of HLT-NAACL* 2003, *Main Papers*, Edmonton, 27 May-1 June 2003, 48-54. https://doi.org/10.3115/1073445.1073462

[25] Michael Carl, M. and Schaeffer, M. (2017) Sketch of a Noisy Channel Model for the Translation Process. In: Hansen-Schirra, S., Czulo, O. and Hofmann, S., Eds., *Empirical Modelling of Translation and Interpreting*, Language Science Press, Berlin, 71-116.

[26] Shannon, C.E. (1948) A Mathematical Theory of Communication. *The Bell System Technical Journal*, **27**, 379-423, 623-656. https://doi.org/10.1002/j.1538-7305.1948.tb00917.x

[27] Miller, G.A. (1955) The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychological Review*, **101**, 343-352. https://doi.org/10.1037/0033-295X.101.2.343

[28] Matricciani, E. (2019) Deep Language Statistics of Italian throughout Seven Centuries of Literature and Empirical Connections with Miller's 7 ± 2 Law and Short-Term Memory. *Open Journal of Statistics*, **9**, 373-406. https://doi.org/10.4236/ojs.2019.93026

[29] Matricciani, E. and De Caro, L. (2019) A Deep-Language Mathematical Analysis of Gospels, Acts and Revelation. *Religions*, **10**, 257.

https://doi.org/10.3390/rel10040257

[30] Papoulis, A. (1990) Probability & Statistics. Prentice Hall, Upper Saddle River.

[31] Matricciani, E. (2009) An Optimum Design of Deep-Space Downlinks Affected by Tropospheric Attenuation. *International Journal of Satellite Communication and Networking*, **27**, 312-329. https://doi.org/10.1002/sat.942

[32] De Caro, L., Giannini, C., Lassandro, R., Scattarella, F., Sibillano, T., Matricciani, E. and Fanti, G. (2019) X-Ray Dating of Ancient Linen Fabrics. *Heritage*, **2**, 2763-2783. https://doi.org/10.3390/heritage2040171

[33] Fano, R.M. (1961) Transmission of Information. A Statistical Theory of Communication. The MIT Press, Cambridge.

[34] Shannon, C.E. (1951) Prediction and Entropy of Printed English. *Bell Labs Technical Journal*, **30**, 50-65. https://doi.org/10.1002/j.1538-7305.1951.tb01366.x

[35] Kandel, L. and Moles, A. (1958) Application de l'indice de Flesch à la langue française. *Cahiers Etudes de Radio- Télévision*, **19**, 253-274.

[36] François, T. (2014) An Analysis of a French as Foreign Language Corpus for Readability Assessment. *Proceedings of the* 3*rd Workshop on NLP for CALL*, *NEALT*, Proceedings 107, Series 22, Linköping, 13-32.

[37] DuBay, W.H. (2004) The Principles of Readability. Impact Information, Costa Mesa.

[38] Lucisano, P. and Piemontese, M.E. (1988) GULPEASE: Una formula per la predizione della difficoltà dei testi in lingua italiana. Scuola e città, 110-124.

[39] Lindgren, B.W. (1968) Statistical Theory. 2nd Edition, MacMillan Company, Basingstoke.

[40] Matricciani, E., Norando, T. and Magnaghi Delfino, P. (2013) How Did Polybius Perfect Cleoxenus and Democritus Signalling Method, 2200 Year Ago? *Technai*, **4**, 33-45.

[41] Manfrino, R. (1960) L'entropia della lingua italiana ed il suo calcolo. *Alta Frequenza*, **29**, 4-29.

[42] Bikesh Revovna, O. (2013) Calculating Information Entropy of Language Texts. *World Applied Sciences Journal*, **22**, 41-45.

Appendix

Appendix A. List of mathematical symbols.

Symbol	Meaning
C	channel capacity
$C_{\max}$	maximum channel capacity
C_P	number of characters per word
ΔC_{kj}	capacity difference
G	readability index
G_c	semantic term
G_F	syntactic term
I	total number of interpunctions
I_P	number of words per interpunction (word interval)
$I_{P\infty}$	horizontal asymptote
m	slope
n_I	number of interpunctions per chapter
n_S	number of sentences per chapter
n_W	number of words per chapter
$m_{C\max}$	slope for maximum capacity C
m_{kj}	slope
M_F	number of interpunctions per sentence
N_m	regression noise power
N_r	correlation noise power
n_k	noise source
$n_{j,tot}$	total noise source
P_F	number of words per sentence
P_{Fo}	$1/e$ of exponential maximum value
r	correlation coefficient
$r_{m=1}$	irreducible correlation coefficient
R	total noise-to-signal power ratio
R_m	regression noise-to-signal power ratio
R_r	correlation noise-to-signal power ratio
$R_{\min}$	minimum noise-to-signal power ratio
ρ	signal-to-noise power ratio
S	total number of sentences
s_x^2	variance of x
s_{nk}^2	noise variance (power)
Y_k	input language
Y_j	output language
W	total number of words

Appendix B. Entropy and human information-processing

The short-term memory capacity follows Miller's 7 ± 2 law [28]. Notice, however, that the range of Miller's law does not refer to bits, but to a "buffer" in which are stored "chunks" of information of the type that can be "compressed", as are sequences of words or sequences of numbers (see [28] and the references there cited). In other words, humans process information differently from translation machines. As a consequence, the entropy of a language may be misleading in studying the linguistic channels defined in this paper. This point is now illustrated with an example.

Let us consider the total number of words W (**Table 1**) of translations into English, French, German, Italian, Russian and Spanish. The entropy of a language referred to single letters is termed F_1 by Shannon [34]. Estimated values of F_1 for the mentioned languages are reported in **Table B1**.

Now the total number of information bits produced according to Communication/Information theory can be estimated by:

$$N_{bit} = W \times C_P \times F_1 \tag{B.1}$$

Table B1 reports the values calculated from Equation (B.1). It is clear that each language/translation has different number of words and bits. **Table B2** reports the ranking of languages (from minimum to maximum) according to the number of words (left column) or the number of bits (right column). The first column is what humans perceive; the second column is what machines process. The two lists are identical only for the first three lines—Russian, Greek and Italian—then they diverge. Now, the short-term memory responds to words not to bits, therefore the use of entropy can be highly misleading (e.g., see German, English, French and Spanish) in estimating quantities and the characteristics of the linguistic channels defined in this paper.

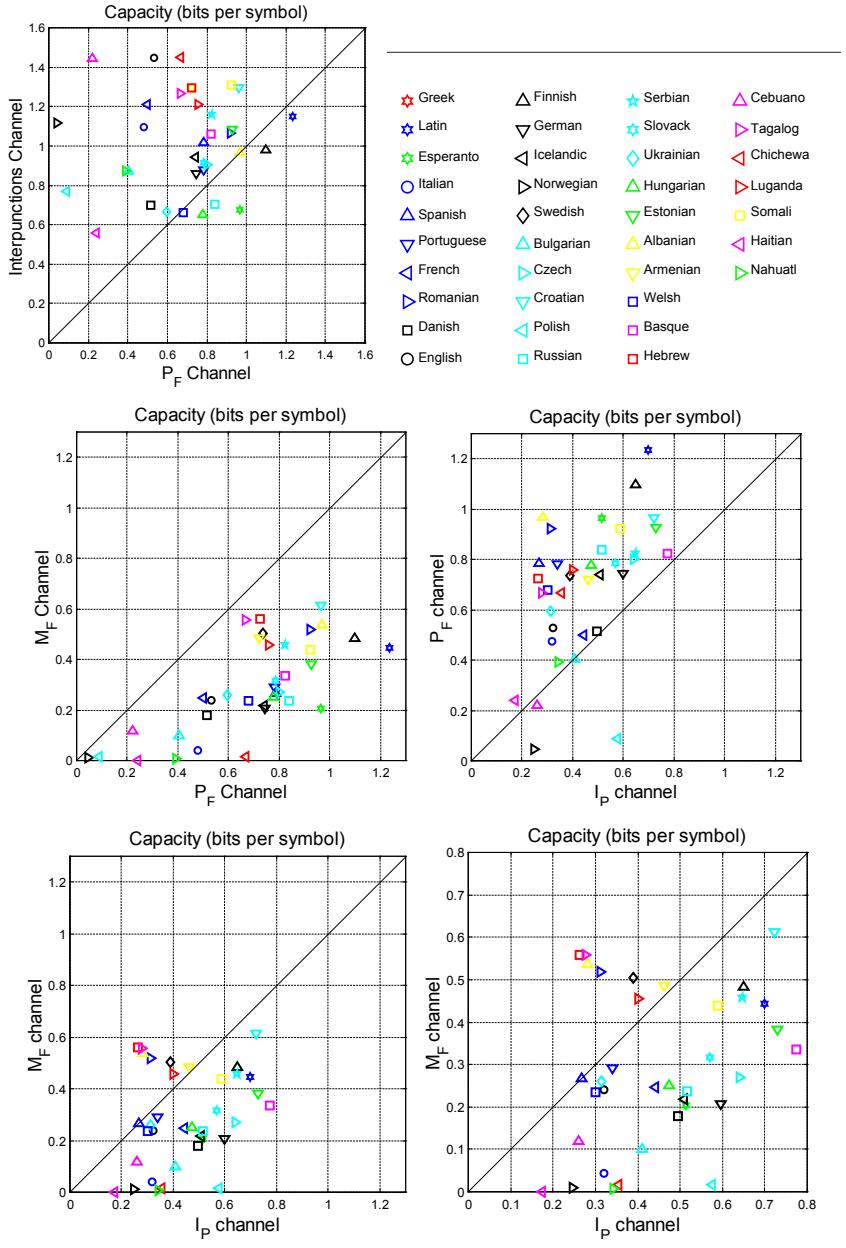
Table B1. Entropy F_1 (bits per letter), total number of words W, average number of letters (characters) per word C_P, difference between the number of bits in the indicated language and in Greek $N_{bit} - N_{Greek}$. Source of F_1 data: [40] for Greek; [41] for French, German, Italian and Spanish; [34] for English; [42] for Russian.

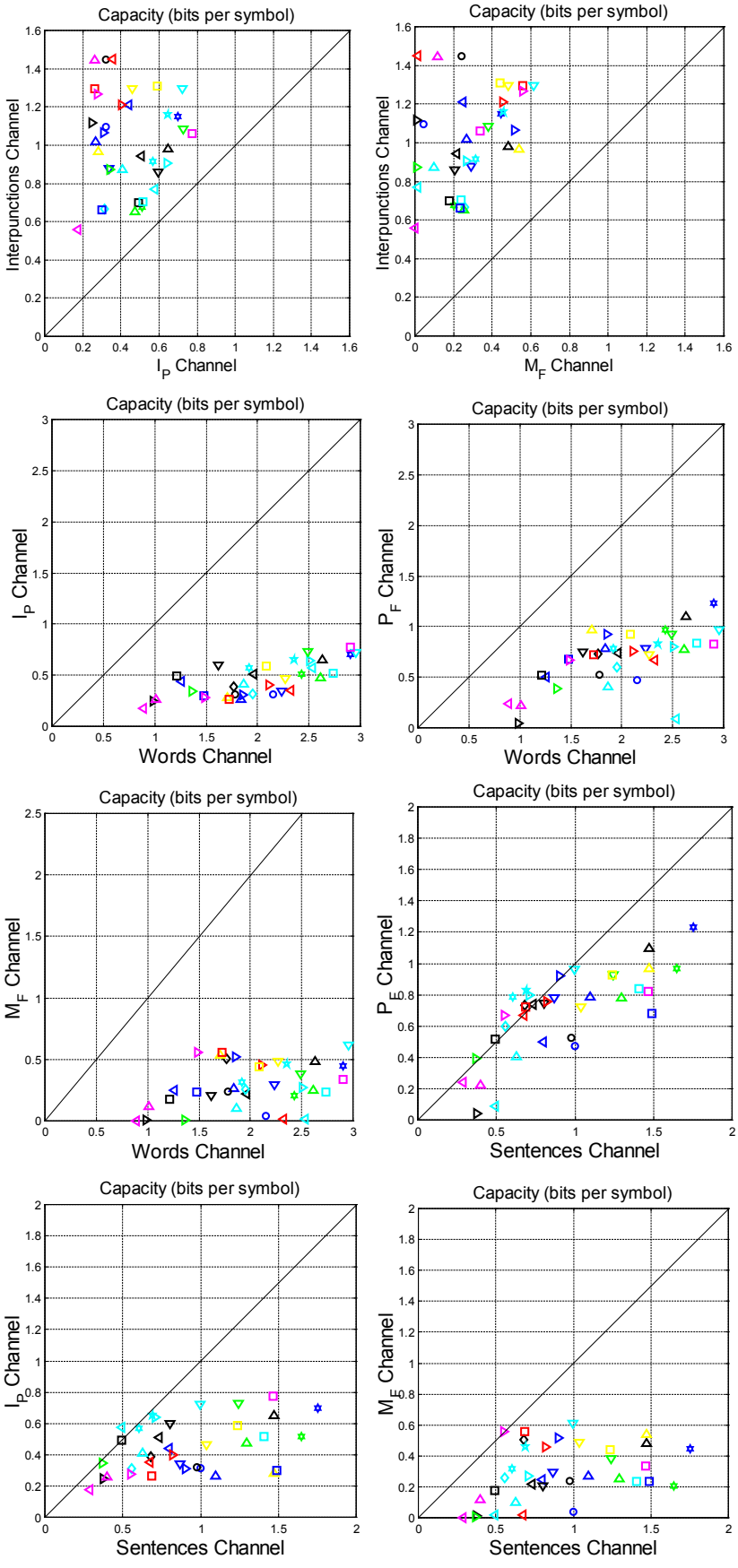
Language	F_1 (bits per letter)	n_W (words)	C_P (letters per word)	N_{bit}	$N_{bit} - N_{bit,Greek}$ (bits)
Greek	4.09	100,145	4.86	1,990,622	0
French	3.98	133,050	4.20	2,224,064	233,442
English	4.14	122,641	4.24	2,152,791	162,169
German	4.08	117,269	4.68	2,239,181	248,559
Italian	3.95	112,943	4.48	1,998,639	8017
Russian	4.36	92,736	4.67	1,888,216	−102,406
Spanish	4.00	118,744	4.30	2,042,397	51,775

Table B2. Ranking (from minimum to maximum).

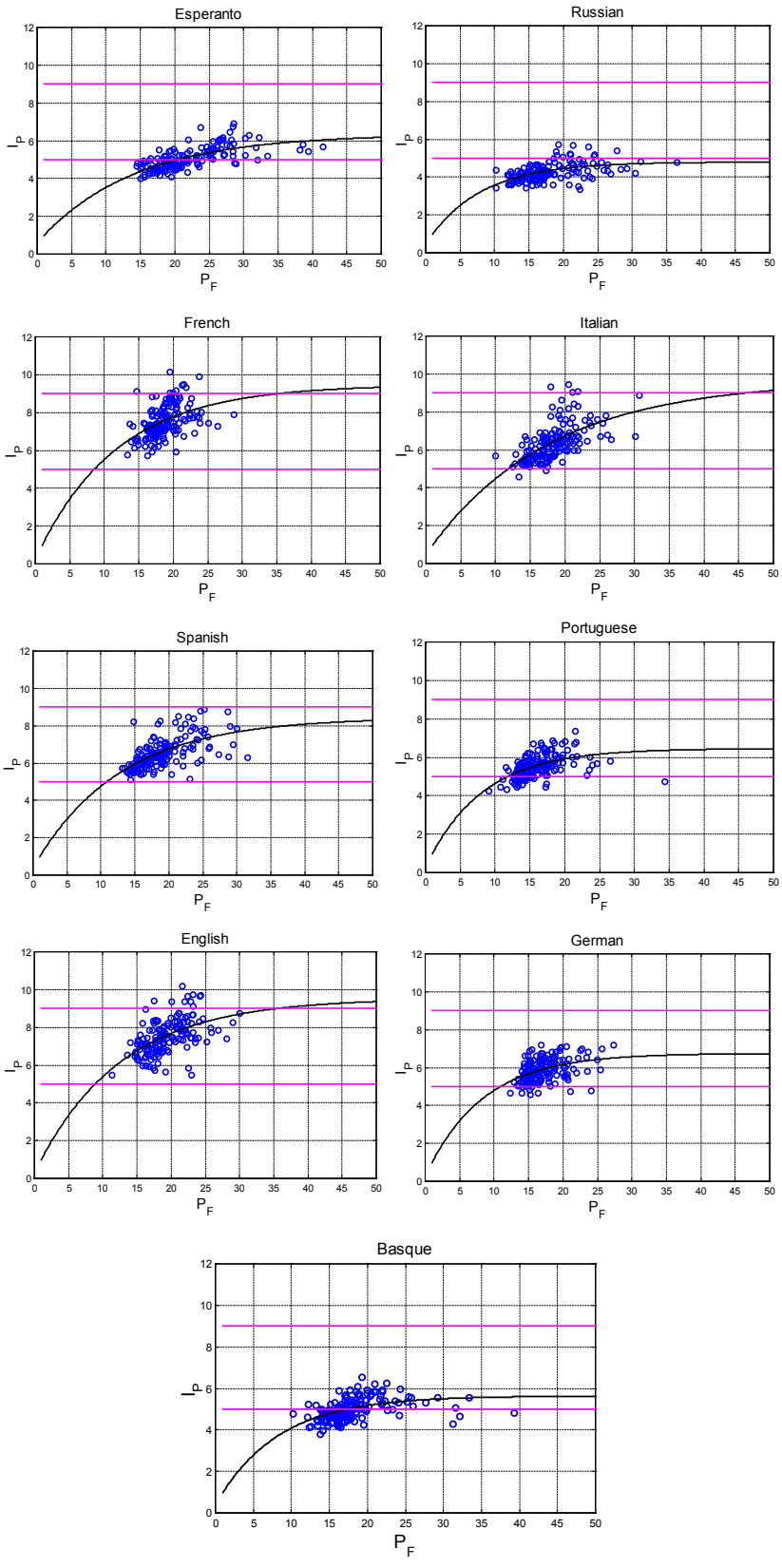
According to words	According to bits
Russian	Russian
Greek	Greek
Italian	Italian
German	Spanish
Spanish	English
English	French
French	German

Appendix C. Scatterplots of average channel capacity for the indicated channels. Languages are distinguished according to the symbols listed below.

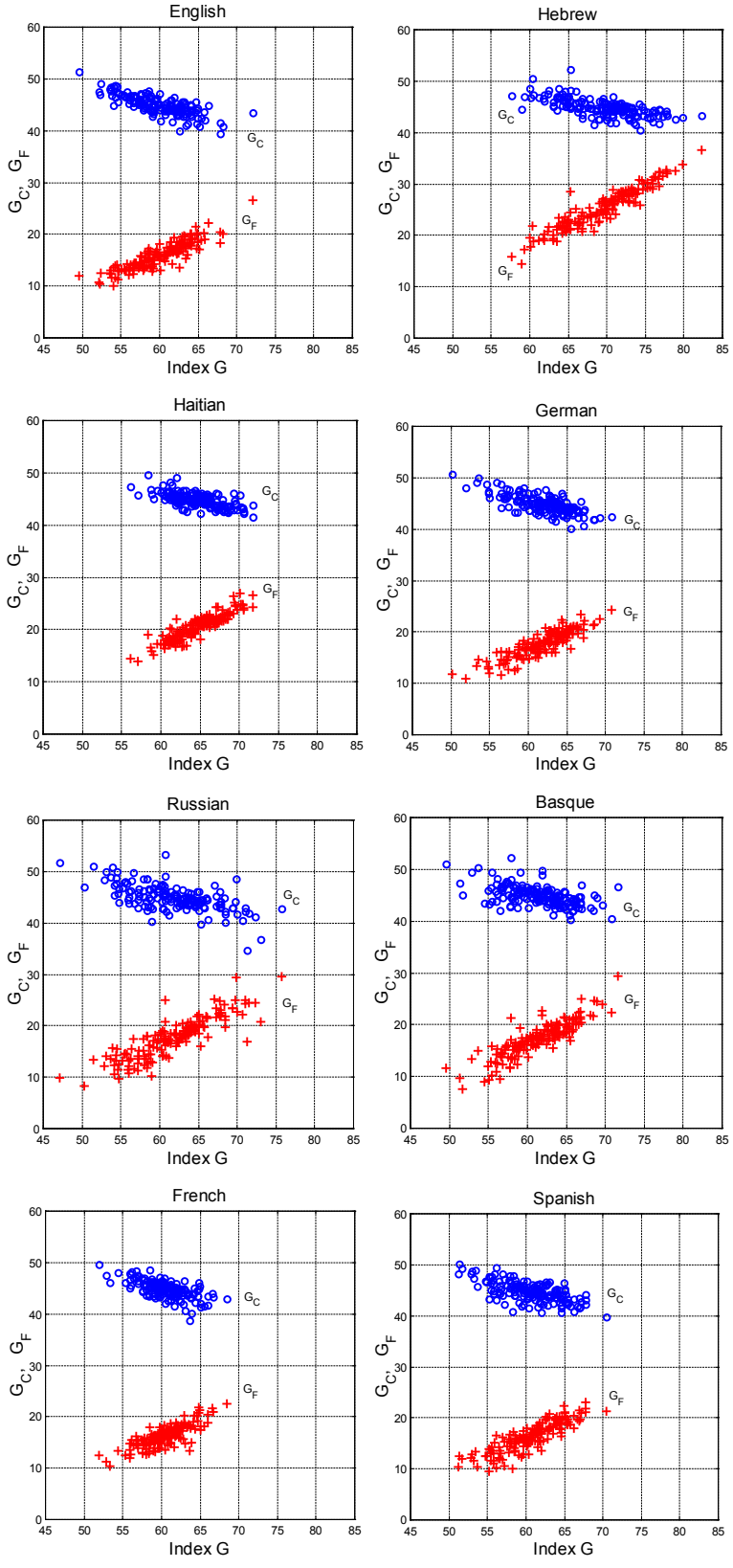


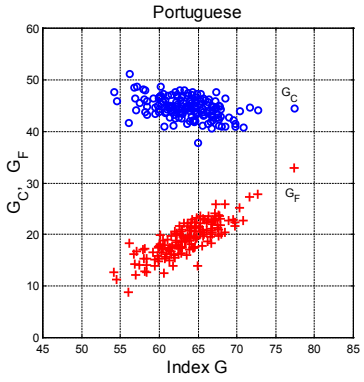

Capacity (bits per symbol)
Interpunctions Channel
I_P Channel
Capacity (bits per symbol)
Interpunctions Channel
M_F Channel
Capacity (bits per symbol)
I_P Channel
Words Channel
Capacity (bits per symbol)
P_F Channel
Words Channel
Capacity (bits per symbol)
M_F Channel
Words Channel
Capacity (bits per symbol)
P_F Channel
Sentences Channel
Capacity (bits per symbol)
I_P Channel
Sentences Channel
Capacity (bits per symbol)
M_F Channel
Sentences Channel

Appendix D. Scatterplots of I_P versus P_F for the indicated languages. The horizontal magenta lines are Miller's bound 5 and 9.



Appendix E. Scatterplots of G_C (blue) and G_F (red) versus G for the indicated languages.





Appendix F. Scatterplots between direct channel capacity (from … to) and the reverse channel capacity (to … from) for all languages. Red circles are the average values of each translation. The red symbol is the overall average value.