# The Small-World of "Le Petit Prince": revisiting the word frequency distribution

D. Gamermann*[1], C. Moret-Tatay[2], E. Navarro-Pardo[3], and P. Fernandez de Córdoba Castellá[4]

[1]Department of Physics, Universidade Federal do Rio Grande do Sul (UFRGS) - Instituto de Física , Av. Bento Gonçalves 9500 - Caixa Postal 15051 - CEP 91501-970 - Porto Alegre, RS, Brasil.
[2]Departamento de Neuropsicobiología, Metodología y Psicología Social - Facultad de Psicología, Magisterio y Ciencias de la Educación, Sede de San Juan Bautista. Universidad Católica de Valencia, San Vicente Mártir - Calle Guillem de Castro 175, 46008- Valencia, Spain.
[3]Department of Developmental and Educational Psychology - Faculty of Psychology, Universitat de València. Av. Blasco Ibáñez, 21 46010 - Valencia, Spain.
[4]Instituto Universitario de Matemática Pura y Aplicada, Universitat Politècnica de València. Camino de Vera, s/n 46022 - Valencia, Spain.

March 6, 2018

**Abstract**

Many complex systems are naturally described through graph theory and different kinds of systems described as networks present certain important characteristics in common. One of these features is the so called scale-free distribution for its node's connectivity, which means that the degree distribution for the network's nodes follows a power law. Scale-free networks are usually refered to as small-world because the average distance between their nodes do not scale linearly with the size of the network, but logarithmically. Here we present a mathematical analysis on linguistics: the word frequency effect for different translations of the "Le Petit Prince" in different languages. Comparison of word association networks with random networks makes evident the discrepancy between the random Erdös-Rény model for graphs and real world networks.

Key words: Small-world, word frequency, Zipf's law

Many objects of study in different interdisciplinary fields find a natural mathematical description as graphs. A graph is simply an object formed by two different sets: a set of nodes and a set of edges connecting these nodes. For many decades the mathematical study of graphs has been guided by the Erdös-Rény model for random graphs Erdös, P. and Rényi, A. (1960). In this model a (random) graph is constructed from a set of $N$ nodes by connecting or not each one of the $\frac{N(N-1)}{2}$ pairs of nodes with a probability $p$. A random graph will, therefore, have on average

*danielg@if.ufrgs.br

$p\frac{N(N-1)}{2}$ links and the degree distribution of its nodes will follow a Poisson distribution. Another characteristic of random graphs is the fact that its size (average node distance) scales linearly with the number of nodes in the graph.

As graph theory started being applied to many real systems such as metabolic or protein networks, neural networks, the Internet, social networks, food-chains, among many others Rives & Galitski (2003), Haykin (1994), Pastor-Satorras et al. (2001), Crucitti et al. (2003), a discrepancy between these real-world graphs and the random Erdös-Rény graphs became evident. The node's degree distribution in real-world graphs do not follow a Poisson distribution, instead they follow a power-law distribution and thus became known as scale-free. As a consequence, the average distance between two nodes in such networks grows slowly with the the number of $N$ nodes in the network and this characteristic is known as small-world behavior Amaral et al. (2000).

It has been observed that the word frequency distribution in a language also follows a scale-free distribution and many explanations for this phenomenon have been given. In linguistics, this observation is known as Zipf's law. It states that the proportion of words $P$ (in a text, for example) with a given frequency $k$ follows a power law: $P(k) \sim k^{-\gamma}$ where $\gamma$ is generally a number between 2 and 3. This law shows that few words present very high frequency and, conversely, many words present low frequency. A particular and appealing explanation for this could be achieved via concepts from statistical mechanics where one tries to minimize an energy function based on the balance between the efforts of the speaker and the listener which is defined by the word frequency and ambiguity, as shown in Cancho & Solé (2003).

One traditional way to examine differences between languages is by variables such as frequency, morphological complexity, evolution and cultural transmission. All these aspects can be related in a complex adaptive system Beckner et al. (2009). In particular, the word frequency is a classical effect in cognitive psychology characterized by its robustness: high frequency words are recognized quicker and remembered better Sternberg & Powell (1983). Therefore, a large body of research has employed the word frequency as an approach of word difficulties Dufau et al. (2011), Esteves et al. (2015), Moreno-Cid et al. (2015), Moret-Tatay & Perea (2011*b*,*a*), Navarro-Pardo et al. (2013), Perea, Moret-Tatay & Carreiras (2011), Perea, Comesaña, Soares & Moret-Tatay (2012), Perea, Gatt, Moret-Tatay & Fabri (2012), Perea, Moret-Tatay & Gómez (2011). According to Breland (1996), the logic of this is that low frequency words are more difficult because they appear less often in print. Moreover, (van Heuven et al. (2014)) proposed the Zipf-scale as a better standardized measure of word frequency. Given the ease with which word counts can be collected at the present time, a useful tool on contrastive linguistics is a lexical corpus of a language. In other words, a large collection of texts in the electronic form supplemented by linguistic annotation that has become an important tool in linguistic studies. Not surprisingly, several databases for Computing Statistics and Psycholinguistic in several languages have been developed for this objective Coltheart (1981), Davis (2005). However, according to Perea et al. (2013), Yap et al. (2011), other variables might be involved in word recognition, in particular in word frequency, such as the number of contexts in which a word appears.

In the present work we focus on the analysis of a single linguistic material (the Little Prince by Saint-Exupéry) in several different languages. To this propose, we have studied statistical properties of the text and networks (graphs) associated with this text. In the different languages we studied the word frequency distribution on one hand and then we constructed different networks by word associations. For each network we built, we evaluated its main properties, like its average clustering coefficient, nodes distances and its degree distribution. In the next section we present the methodology we used and the mathematics behind our analyses, in the Results section we describe our findings and in the Conclusions section we present the main aspects of our results and a brief overview.

# 1 Methods

## 1.1 Materials

The Little Prince text was obtained from the Internet in eight different languages: Spanish, English, Dutch, Greek, Basque, Italian, Portuguese and (of course) French.

In order to analyze the text, python scripts were written. The computer codes were run in a computer with a i7 quadcore processor and 8Gb of RAM memory. The scripts first stored all text in the computer RAM memory. Then, it used punctuation in order to slice the text in its sentences and then removed all punctuation and numerals (0, 1, 2, ...) from the raw text. It then identified the different words as the strings left which were separated by spaces. As an example, below one can see the first 300 characters from the French text:

Antoine de Saint-Exupéry

LE PETIT PRINCE

1943

PREMIER CHAPITRE

Lorsque j'avais six ans j'ai vu, une fois, une magnifique image, dans un livre sur la Forêt Vierge qui s'appelait « Histoires Vécues ». Ça représentait un serpent boa qui avalait un fauve. Voilà la copie du dessin. On disai

Through our scripts, the extract above becomes the list of words: antoine, de, saint, exupéry, le, petit, prince, premier, chapitre, lorsque, j, avais, six, ans, j, ai, vu, une, fois, une, magnifiqu image, dans, un, livre, sur, la, forêt, vierge, qui, s, appelait, histoires, vecues, ça, représenta un, serpent, boa, qui, avalait, un, fauve, voilà, la, copie, du, dessin, on, disait.

Once the python script transforms the whole text in a raw list of words (15612 in the case of the French text), it counts the number of different words (2600 in the French text) and counts also the number of times that each single word is repeated in the text. For the construction of networks, we will link words based on their relative distance in the text. For this, one needs to keep track of the sentences in which the text is divided and which words appear in each sentence. So our script actually first creates a list of sentences, by slicing the text when it finds a punctuation symbol, and after that a list of single words, by slicing the sentences in its blank spaces.

## 1.2 Analysis

The word frequency distribution $P(k)$ is a function that, for each natural number $k$, tells how many words appeared in the text $k$ times. In the case of the French text, for example, 1516 different words appeared only once ($P(1) = 1516$), one of these is the word "réjouir", that appears in the whole text only once. On the other hand, the word "et" was the fifth most frequent word, appearing 306 times ($k = 306$) and this is the only word that appeared this number of times, consequently $P(306) = 1$. The most frequent word was the article "`le`" that appeared 465 times and is the only word appearing 465 times in the text ($P(465) = 1$).

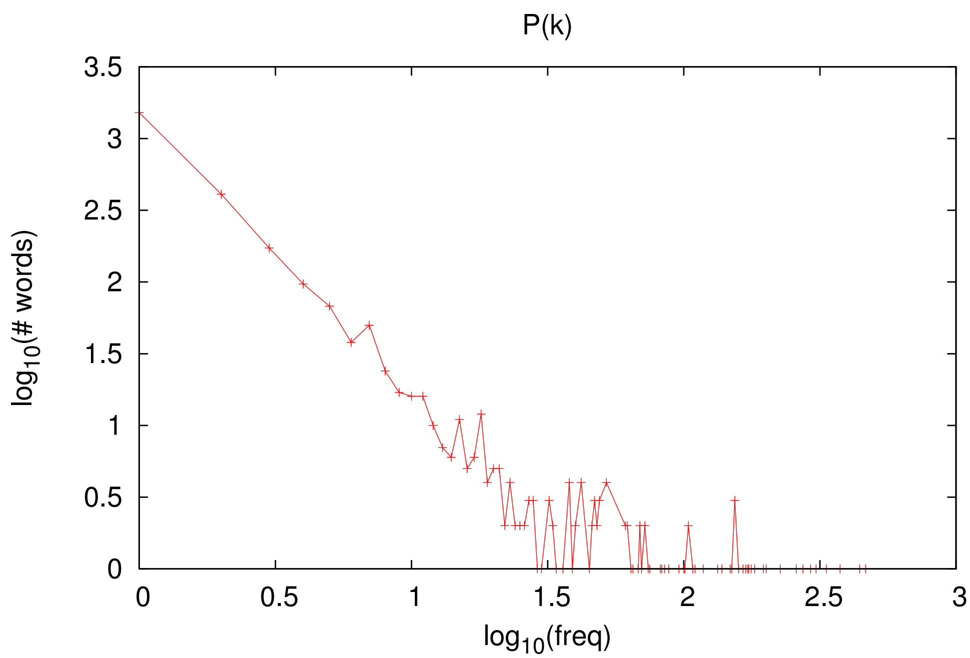
Typically, for a text, many words appear only a few times, while a few words are repeated constantly along the text. As a consequence, the function $P(k)$ is a decreasing function. A mathematical function that often fits $P(k)$ in a text is the power-law distribution:

$$P(k) = Ak^{-\gamma}, \quad (1)$$
$$\log(P(k)) = \log(A) - \gamma \log(k) \quad (2)$$

where $A$ is a proportionality constant that can be evaluated by the total number of words. The fact that the frequency distribution follows a power-law (or scale-free) distribution is known as the Zipf law. Note from equation (2) that, in a log-log plot, the distribution will follow a straight line.

For real texts, the tail (large values of $k$) of the $P(k)$ distribution will be very noisy, because only a handful of large values of $k$ will be populated and then by a single word. In figure 1 we show the function $P(k)$ (in logarithmic scale) for the French text. One can clearly see the noise in the right tail.

Figure 1: Word frequency distribution for the French text with a noisy right tail.



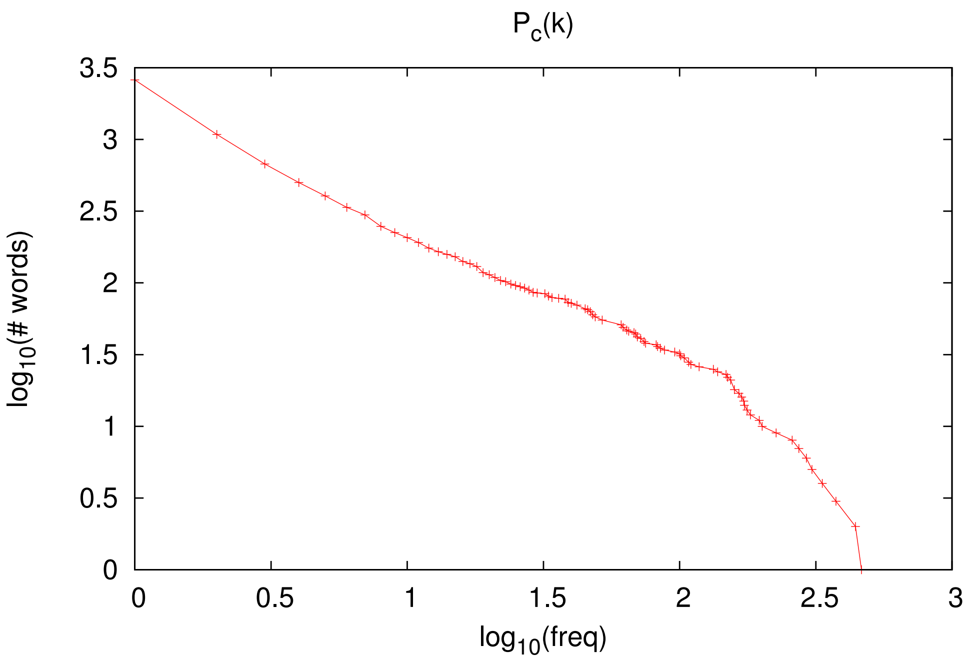
In order to fit the distribution avoiding the noisy tail, one can use the right-cumulative distribution:

$$P_c(k) = \int_k^\infty P(k')dk' = \frac{A}{\gamma - 1} k^{-(\gamma-1)} \quad (3)$$

$$\log\left(P_c(k)\right) = \log\left(\frac{A}{\gamma - 1}\right) - (\gamma - 1)\log(k). \quad (4)$$

In figure 2 one can see the distribution $P_c(k)$ (in logarithmic scale) for the French text. This curve is much smoother than the raw $P(k)$ distribution and it is always decreasing.

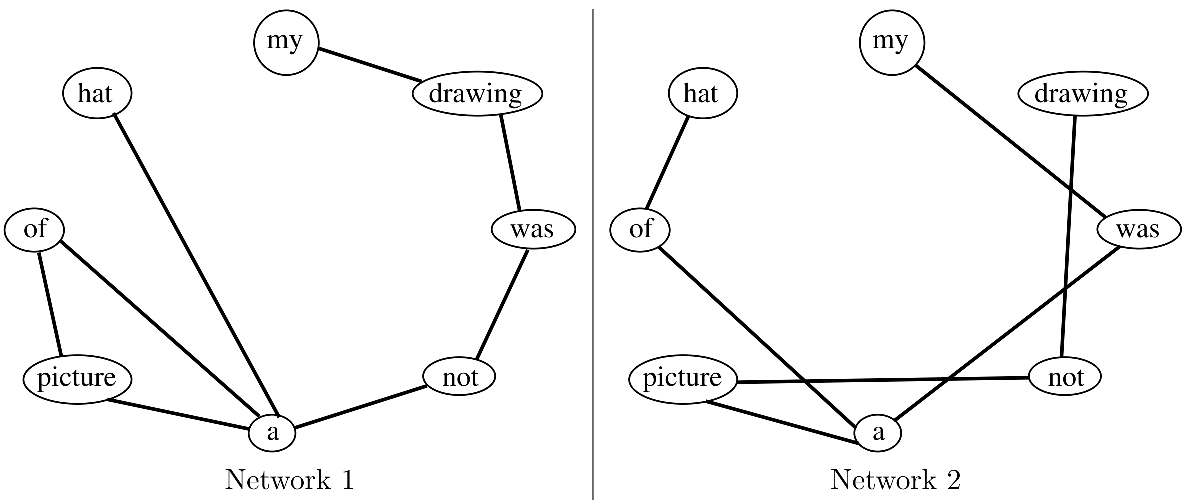
Figure 2: Word frequency cumulative distribution for the French text.




From equations (2) and (4) it is clear that the plot of $\log(P)$ or $\log(P_c)$ versus $\log(k)$ will follow a straight line if the distribution $P(k)$ follows the power-law in equation (1). So fitting lines to the empirical data collected from the texts, one can determine the parameters $A$ and $\gamma$. The parameter $A$ divided by $\gamma - 1$ is just the total number of different words in a text. One can realize this by noticing that $P_c(1) = \#\text{total of words}$.

Apart from measuring and fitting the word frequency distribution, we analyzed networks of words association built from the texts. In order to build a network from the texts in each language, we set each word as a node and we built two different networks by following two different rules in order to set the links between words. In the first network we define a link between two words if they appear side by side in at least one sentence in the text. In the second network a link is defined between two words if there is a third word between the two in at least one sentence in the text. In figure 3 we show examples of the two networks based on a single sentence in the text: "My drawing was not a picture of a hat!"

Figure 3: Example of the two networks. Network 1 on the left and network 2 on the right.




An important structure in order to analyze a graph is its adjacency matrix, this is a symmetric $N \times N$ matrix, where $N$ is the number of nodes in the graph and the elements $M_{ij}$ are equal to one if there is a link between nodes $i$ and $j$ and zero otherwise. From this matrix, one can directly obtain the degree (number of neighbors or connections) for any given node in the graph: $k_i = \sum_{j=1}^{N} M_{ij}$.

The number of nodes (words) in each network constructed from the texts maybe less than the total number of different words in each whole text because we remove non-connected components (sets of nodes from which it is not possible the reach a bigger set of nodes following the links within the set) from the graphs. For each network we performed three analyzes: we fitted a power-law to its degree distribution, we calculated the average clustering coefficient and the average distance between two nodes.

The fitting of a power-law follows the same steps done in order to fit word frequencies (but now looking at degree for each node in the network). The clustering coefficient of a node is given by Ravasz & Barabasi (2003):

$$C_i = \frac{2E_i}{k_i(k_i-1)} \tag{5}$$

where $k_i$ is the degree of node $i$ and $E_i$ is the number of connections between the neighbors of node $i$. The average clustering $\bar{C}$ of a network can now be calculated straightforward as the average value of the $C_i$'s for all nodes in the network.

The distance between two nodes is defined as the minimum number of links one has to go through in order to travel from one node to the other. The average distance between every one of the $\frac{N(N-1)}{2}$ different pairs of nodes in each network was calculated using Dijkstra's algorithm Dijkstra (1959) via the PyNetMet package Gamermann et al. (2014). The average of the distances between every pair is the network's average distance $\bar{d}$.

We compared the average clustering and average distance in every network with results from random networks. For this purpose, for each network, we built an ensemble with twenty random networks with the same number of nodes and the same number of links, but with random topology. The input for a network is its adjacency matrix $M$. So, for building a random network we use the following algorithm:

(1) Start with an $N \times N$ matrix where all its elements are zero. (One has here N nodes and zero links ($\ell = 0$) between them.

(2) While the number of links ($\ell$) is less than the desired number of links in the network, repeat:

  (2.1) Chose two different integer random numbers ($i$ and $j$) between 1 and $N$.

  (2.2) If $M_{ij}$ is zero, change $M_{ij}$ and $M_{ji}$ to one and increase in one unit the number of links ($\ell \rightarrow \ell + 1$).

(3) Check if any node ($i$) has been left unconnected. If so, randomly choose a node ($j$) to connect it ($i$) to and randomly break an existing connection of node $j$.

(4) Repeat step (3) until no node is left unconnected.

Steps (3) and (4) are actually optional, but throughout our calculations, we have chosen to work with fully connected graphs. This algorithm returns a randomly generated adjacency matrix representing a connected network with a predefined number of nodes and links.

Using this algorithm, for each network obtained from a text, we generate an ensemble of twenty random networks with the same number of nodes and links. For each random network in the ensemble the average clustering and average distance is calculated and then the average inside each ensemble is evaluated.

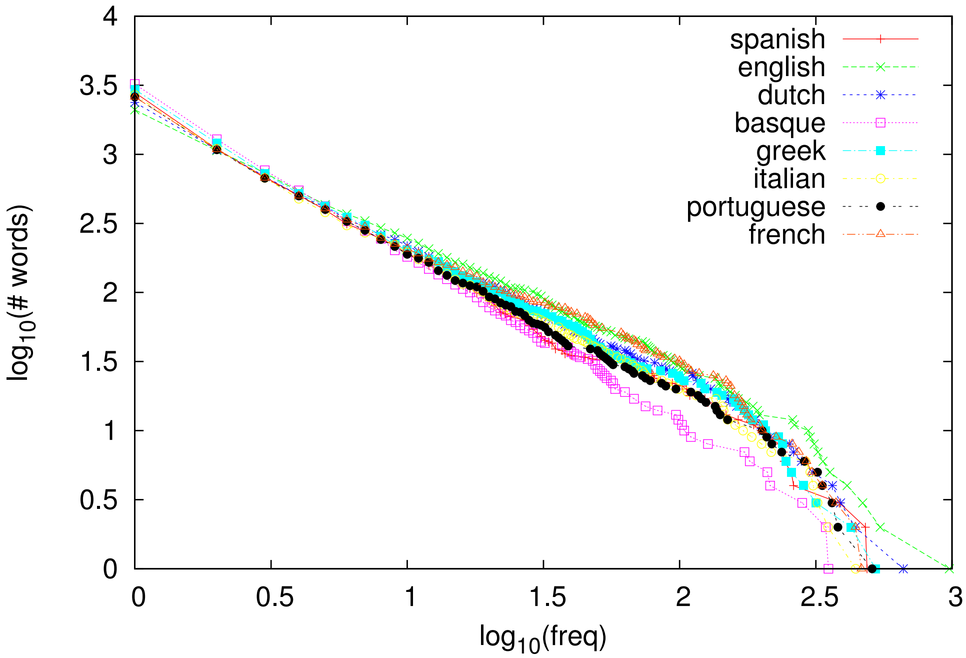
## 2 Results

In figure 4 the distributions for all the eight languages in log-log scale are supper-posed showing the tendency they have to follow a straight line. In figure 5 the distribution for each individual language is shown with the best line fitted using the least squares method. In the title of each plot one finds the equation fitted.

In table 1 we show the values of $\gamma$, $\frac{A}{\gamma-1}$, total number of words and the $\frac{\chi^2}{dof}$ for the best fit for each language. The value for $\chi^2$ (minimized by the least square method) is calculated as:

$$\chi^2 = \sum_{k=1}^{k_{max}} \frac{(\log(P_c(k)) - \log(P_{c_{obs,k}}))^2}{\epsilon_k} \tag{6}$$

Figure 4: Cumulative word frequency distribution for all texts.




where $P_{c_{obs,k}}$ is the observed value for the right-cumulative distribution of words at frequency $k$, $\epsilon_k$ is the error associated to $\log(P_{c_{obs,k}})$ and the sum is made for all $k$'s for which $P_{obs,k}$ is different from zero [1]. Since $P_{c_{obs,k}}$ is an absolute frequency, the error associated to it is its square-root and, therefore, one evaluates the logarithmic[2] error $\epsilon_k = \frac{1}{\ln(10)\sqrt{P_{c_{obs,k}}}}$.

The results for the networks analysis can be found in tables 2 and 3. In figure 6 we show, for the Network 1 constructed from the Portuguese text, its degree distribution, the best fitted line to it and the degree distribution for a random network with the same number of nodes and links ($N = 2424$ and $\ell = 6175$). From this figure, one can clearly see the difference between the distribution obtained from a "real" network (power-law distribution) and the one obtained from a completely random network (Poisson distribution). In a power-law distribution there is a sensible probability of observing nodes with a higher (much bigger than average) degree, while in a Poisson distribution this probability drops to zero very fast.

[1]Note that $P_{c_{obs,k}}$ is the right cumulative distribution so, if $P_{obs,k}$ is zero for a given value of $k$, $P_{c_{obs,k}}$ will be a constant for all $k's$ after this, until reaching a new $k$ where $P_{obs,k}$ is not zero, and therefore, these points would not bring any new information to the analysis.

[2]In all our equations log is the base 10 logarithm and ln is the natural (base $e$) logarithm.

Figure 5: Cumulative word frequency distribution for all texts with the best line fitted.

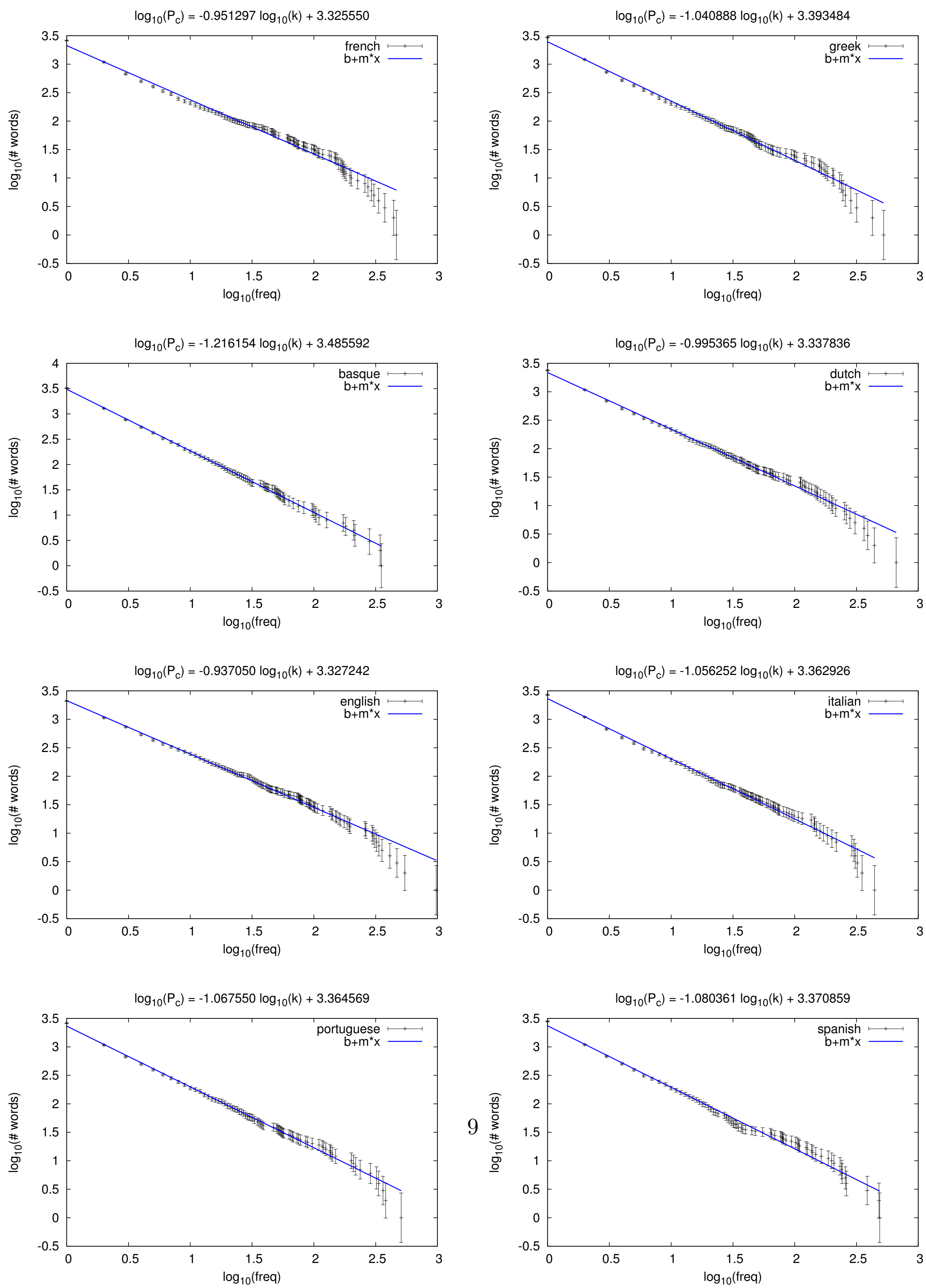


Figure 6: Degree distribution for the Network 1 obtained from the Portuguese text compared with a random network.

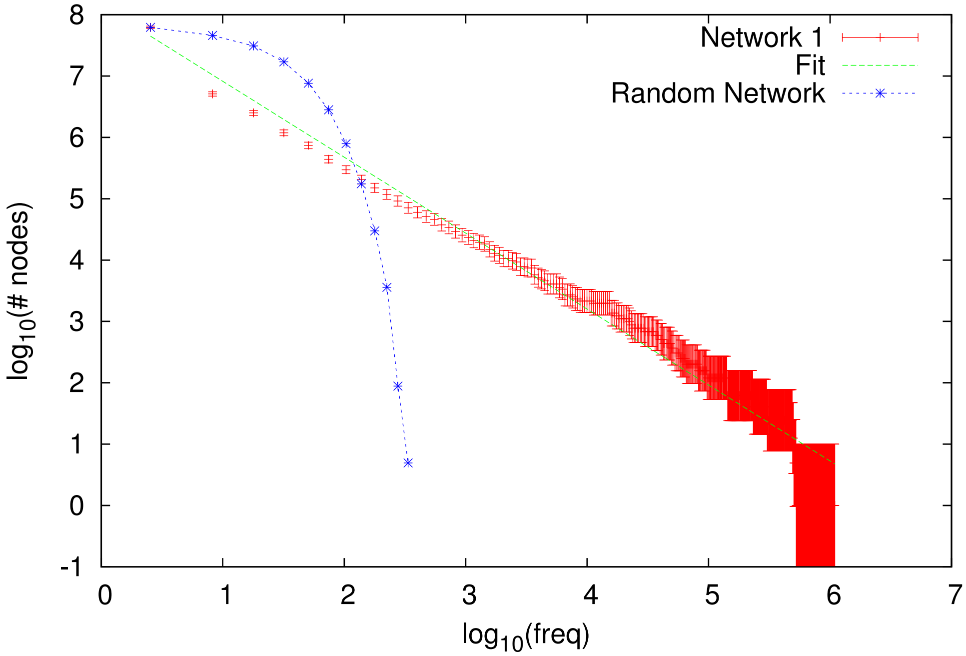


Table 1: Summary of the fits.

| Language | # words | $\frac{A}{\gamma-1}$ | $\gamma$ | $\frac{\chi^2}{dof}$ |
|---|---|---|---|---|
| SPANISH | 2801 | 2348.87 | 2.08 | 0.078 |
| ENGLISH | 2098 | 2124.43 | 1.94 | 0.041 |
| DUTCH | 2375 | 2176.89 | 2.00 | 0.040 |
| BASQUE | 3226 | 3059.09 | 2.22 | 0.016 |
| GREEK | 2951 | 2474.48 | 2.04 | 0.063 |
| ITALIAN | 2689 | 2306.35 | 2.06 | 0.045 |
| PORTUGUESE | 2607 | 2315.10 | 2.07 | 0.031 |
| FRENCH | 2600 | 2116.17 | 1.95 | 0.112 |

The properties calculated for the two types of networks (1 and 2) are very similar, but they differ significantly from the properties calculated for random networks. The average node distance in the random networks are, on average, around two units bigger than in the language networks and they present a much smaller standard deviation in the case of random networks. The second interesting difference between random and language networks is the average clustering coefficient, which is very close to zero in the case of random networks. In language networks, words tend to form clusters because of the language structure (they will share either context, grammatical or semantic function, ...) and this feature is reflected in the clustering coefficient calculated from eq. (5).

## 3 Conclusions

Here we present a mathematical analysis on linguistics: the word frequency effect for different translations of the same book ("Le Petit Prince") in eight different languages. The interest of these studies is that the occurrence of words in sentences reflects the language's organization. Apart from the word frequency distribution, we also performed analyzes of different networks built based on word associations in the text and compared these to random networks.

As expected, word frequency presented a scaling law. The results suggest small differences on language volume for the same material. In particular, the $\gamma$ parameter varied slightly across the different languages. Moreover, our study shows how different languages tend to slightly differ in formal aspects. Comparison of word association networks with random networks makes evident the discrepancy between the random Erdös-Rény model for graphs and real world networks. A real network follows a specific design principle and therefore its nodes are connected in an organized way. This becomes evident from the clustering coefficient of the networks which have a high value for networks 1 and 2, but is very close to zero for the random networks. Another interesting difference between the real and random networks is the observation of the small-world effect in real networks: its average node's distance is much smaller than in random networks.

Finally, one can conclude that these results show how different languages tend to slightly differ in formal aspects

Table 2: Network 1 parameters for the different languages. $N$ is the number of nodes and $\ell$ is the number of links, $\gamma$ is the parameter obtained fitting a power-law to the degree distribution for the nodes, $\bar{C}$ is the average clustering, $\bar{d}$ is the average nodes distances. The parameters with a subscript $R$ refer to the the averages in the random networks and the uncertainties shown are the standard deviations for the calculated averages (in the case $\bar{C}_R$ and $\bar{d}_R$, it is the standard deviation within the ensemble and not the average standard deviation within networks).

| Language | $N$ | $\ell$ | $\gamma$ | $\bar{C}$ | $\bar{d}$ | $\bar{C}_R$ | $\bar{d}_R$ |
|---|---|---|---|---|---|---|---|
| SPANISH | 2705 | 6912 | 2.223 | 0.203 ± 0.343 | 3.240 ± 0.416 | 0.002 ± 0.000 | 4.988 ± 0.015 |
| ENGLISH | 1950 | 6770 | 2.260 | 0.248 ± 0.358 | 3.026 ± 0.379 | 0.004 ± 0.001 | 4.123 ± 0.006 |
| DUTCH | 2236 | 7048 | 2.201 | 0.294 ± 0.440 | 3.156 ± 0.413 | 0.003 ± 0.001 | 4.388 ± 0.006 |
| BASQUE | 3100 | 7017 | 2.481 | 0.069 ± 0.219 | 3.915 ± 0.657 | 0.001 ± 0.000 | 5.408 ± 0.021 |
| GREEK | 2745 | 6990 | 2.273 | 0.210 ± 0.349 | 3.287 ± 0.494 | 0.002 ± 0.000 | 5.005 ± 0.013 |
| ITALIAN | 2559 | 6566 | 2.258 | 0.153 ± 0.302 | 3.363 ± 0.446 | 0.002 ± 0.000 | 4.946 ± 0.014 |
| PORTUGUESE | 2311 | 5786 | 2.240 | 0.198 ± 0.365 | 3.292 ± 0.442 | 0.002 ± 0.000 | 4.945 ± 0.020 |
| FRENCH | 2230 | 6004 | 2.327 | 0.207 ± 0.362 | 3.231 ± 0.391 | 0.002 ± 0.001 | 4.737 ± 0.017 |

when the context is controlled. In particular, these results are of interest to other applied fields. Bear in mind that, in recent decades, the cognitive psychology has paid particular interest to examining factors influencing the recognition of printed words, i.e., frequency, familiarity, word length, age of acquisition among others, according to Andrews (2006). There remain some empirical underlying questions, regarding the question of measuring the word frequency for different languages, from printed manuals to even subtitles. Even if more research is needed here, the comparison between these sources is beyond the scope of this study. Here, we offer a comparison employing different translations of the same printed material in different languages. That allows us to compare differences of word frequency in the same context. Regarding this topic, Perea et al. (2013), Yap et al. (2011) stated that other variables must have a role on frequency, such as the number of contexts in which a word appears. That correspond with the nature of our results. Furthermore, some researchers (van Heuven et al. (2014)) proposed the Zipf-scale as a better standardized measure of word frequency, giving also examples of printed words with various Zipf values. The authors also claimed that an alternative Zipf scale presented in their work is better suited for research in word recognition. Here, we follow the same logic. Thus, these results might offer some insights in to the role of the word frequency effect for print words, but more research in this field is necessary.

## Acknowledgment

We would like to thank Thomas Irvin for his invaluable help and comments.

Table 3: Network 2 parameters for the different languages. $N$ is the number of nodes and $\ell$ is the number of links, $\gamma$ is the parameter obtained fitting a power-law to the degree distribution for the nodes, $\bar{C}$ is the average clustering, $\bar{d}$ is the average nodes distances. The parameters with a subscript $R$ refer to the the averages in the random networks and the uncertainties shown are the standard deviations for the calculated averages (in the case $\bar{C}_R$ and $\bar{d}_R$, it is the standard deviation within the ensemble and not the average standard deviation within networks).

| Language | $N$ | $\ell$ | $\gamma$ | $\bar{C}$ | $\bar{d}$ | $\bar{C}_R$ | $\bar{d}_R$ |
|---|---|---|---|---|---|---|---|
| SPANISH | 2682 | 6418 | 2.233 | 0.262 ± 0.518 | 3.413 ± 0.644 | 0.002 ± 0.001 | 5.164 ± 0.017 |
| ENGLISH | 1927 | 6499 | 2.277 | 0.332 ± 0.513 | 3.129 ± 0.506 | 0.003 ± 0.000 | 4.167 ± 0.009 |
| DUTCH | 2218 | 6577 | 2.213 | 0.370 ± 0.611 | 3.145 ± 0.560 | 0.003 ± 0.001 | 4.515 ± 0.010 |
| BASQUE | 3035 | 6064 | 2.439 | 0.157 ± 0.416 | 3.792 ± 0.948 | 0.001 ± 0.000 | 5.784 ± 0.024 |
| GREEK | 2703 | 6266 | 2.321 | 0.221 ± 0.481 | 3.438 ± 0.803 | 0.002 ± 0.000 | 5.250 ± 0.018 |
| ITALIAN | 2537 | 6203 | 2.283 | 0.163 ± 0.367 | 3.478 ± 0.654 | 0.002 ± 0.001 | 5.068 ± 0.019 |
| PORTUGUESE | 2260 | 5064 | 2.285 | 0.232 ± 0.476 | 3.425 ± 0.792 | 0.002 ± 0.000 | 5.230 ± 0.016 |
| FRENCH | 2191 | 5290 | 2.298 | 0.202 ± 0.447 | 3.366 ± 0.712 | 0.002 ± 0.001 | 5.007 ± 0.015 |

# References


Amaral, L. A., Scala, A., Barthelemy, M. & Stanley, H. E. (2000), 'Classes of small-world networks', *Proc. Natl. Acad. Sci. U.S.A.* **97**(21), 11149–11152.

Andrews, S. (2006), 'All about words: A lexicalist perspective on reading', *From inkmarks to ideas: Current issues in lexical processing* p. 318.

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D. & Schoenemann, T. (2009), 'Language is a complex adaptive system: Position paper', *Language learning* **59**(s1), 1–26.

Breland, H. M. (1996), 'Word frequency and word difficulty: A comparison of counts in four corpora', *Psychological Science-Cambridge-* **7**, 96–99.

Cancho, R. F. & Solé, R. V. (2003), 'Least effort and the origins of scaling in human language', *Proceedings of the National Academy of Sciences* **100**(3), 788–791.

Coltheart, M. (1981), 'The mrc psycholinguistic database', *The Quarterly Journal of Experimental Psychology* **33**(4), 497–505.

Crucitti, P., Latora, V., Marchiori, M. & Rapisarda, A. (2003), 'Efficiency of scale-free networks: error and attack tolerance', *Physica A: Statistical Mechanics and its Applications* **320**, 622–642.

Davis, C. J. (2005), 'N-watch: A program for deriving neighborhood size and other psycholinguistic statistics', *Behavior research methods* **37**(1), 65–70.

Dijkstra, E. (1959), 'A note on two problems in connexion with graphs', *Numerische Mathematik* **1**(1), 269–271. **URL:** *http://dx.doi.org/10.1007/BF01386390*

Dufau, S., Duñabeitia, J. A., Moret-Tatay, C., McGonigal, A., Peeters, D., Alario, F.-X., Balota, D. A., Brysbaert, M., Carreiras, M., Ferrand, L. et al. (2011), 'Smart phone, smart science: how the use of smartphones can revolutionize research in cognitive science', *PloS one* **6**(9), e24974.

Erdös, P. and Rényi, A. (1960), On the Evolution of Random Graphs, *in* 'Publication of the mathematical institute of the hungarian academy of sciences', , pp. 17–61.

Esteves, C. S., Oliveira, C. R., Moret-Tatay, C., Navarro-Pardo, E., Carli, G. A. D., Silva, I. G., Irigaray, T. Q. & Argimon, I. I. d. L. (2015), 'Phonemic and semantic verbal fluency tasks: normative data for elderly brazilians', *Psicologia: Reflexão e Crítica* **28**(2), 350–355.

Gamermann, D., Montagud, A., Jaime Infante, R., Triana, J., Urchueguía, J. & Fernández de Córdoba, P. (2014), 'Pynetmet: Python tools for efficient work with networks and metabolic models', *Computational and Mathematical Biology* (3), 1–11.

Haykin, S. (1994), *Neural Networks: A Comprehensive Foundation*, 1st edn, Prentice Hall PTR, Upper Saddle River, NJ, USA.

Moreno-Cid, A., Moret-Tatay, C., Irigaray, T. Q., Argimon, I. I., Murphy, M., Szczerbinski, M., Martínez-Rubio, D., Beneyto-Arrojo, M. J., Navarro-Pardo, E. & Fernández, P. (2015), 'The role of age and emotional valence in word recognition: An ex-gaussian analysis', *Studia Psychologica* **57**(2), 83–94.

Moret-Tatay, C. & Perea, M. (2011*a*), 'Do serifs provide an advantage in the recognition of written words?', *Journal of Cognitive Psychology* **23**(5), 619–624.

Moret-Tatay, C. & Perea, M. (2011*b*), 'Is the go/no-go lexical decision task preferable to the yes/no task with developing readers?', *Journal of experimental child psychology* **110**(1), 125–132.

Navarro-Pardo, E., Navarro-Prados, A. B., Gamermann, D. & Moret-Tatay, C. (2013), 'Differences between young and old university students on a lexical decision task: Evidence through an ex-gaussian approach', *The Journal of General Psychology* **140**(4), 251–268.

Pastor-Satorras, R., Vazquez, A. & Vespignani, A. (2001), 'Dynamical and correlation properties of the internet', *Phys. Rev. Lett.* **87**(25), 258701.

Perea, M., Comesaña, M., Soares, A. P. & Moret-Tatay, C. (2012), 'On the role of the upper part of words in lexical access: Evidence with masked priming', *The Quarterly Journal of Experimental Psychology* **65**(5), 911–925.

Perea, M., Gatt, A., Moret-Tatay, C. & Fabri, R. (2012), 'Are all semitic languages immune to letter transpositions? the case of maltese', *Psychonomic bulletin & review* **19**(5), 942–947.

Perea, M., Moret-Tatay, C. & Carreiras, M. (2011), 'Facilitation versus inhibition in the masked priming same–different matching task', *The Quarterly Journal of Experimental Psychology* **64**(10), 2065–2079.

Perea, M., Moret-Tatay, C. & Gómez, P. (2011), 'The effects of interletter spacing in visual-word recognition', *Acta psychologica* **137**(3), 345–351.

Perea, M., Soares, A. P. & Comesaña, M. (2013), 'Contextual diversity is a main determinant of word identification times in young readers', *Journal of experimental child psychology* **116**(1), 37–44.

Ravasz, E. & Barabasi, A. L. (2003), 'Hierarchical organization in complex networks', *Phys Rev E Stat Nonlin Soft Matter Phys* **67**(2 Pt 2), 026112.

Rives, A. W. & Galitski, T. (2003), 'Modular organization of cellular networks', *Proc. Natl. Acad. Sci. U.S.A.* **100**(3), 1128–1133.

Sternberg, R. J. & Powell, J. S. (1983), 'Comprehending verbal comprehension.', *American Psychologist* **38**(8), 878.

van Heuven, W. J., Mandera, P., Keuleers, E. & Brysbaert, M. (2014), 'Subtlex-uk: A new and improved word frequency database for british english', *The Quarterly Journal of Experimental Psychology* **67**(6), 1176–1190.

Yap, M. J., Tan, S. E., Pexman, P. M. & Hargreaves, I. S. (2011), 'Is more always better? effects of semantic richness on lexical decision, speeded pronunciation, and semantic classification', *Psychonomic Bulletin & Review* **18**(4), 742–750.