Alma Mater Studiorum · Università di Bologna

FACOLTÀ DI SCIENZE MATEMATICHE, FISICHE E NATURALI
Corso di Laurea Magistrale in Matematica

# STATISTICAL ANALYSIS OF WRITTEN LANGUAGES

**Tesi di Laurea in Sistemi Dinamici e Applicazioni**

**Relatore:**
**Chiar.mo Prof.**
**Mirko Degli Esposti**

**Presentata da:**
**Tommaso Pola**

**Correlatore:**
**Dott.**
**Giampaolo Cristadoro**

**Sessione II**
**Anno Accademico 2012/2013**

*To Metrey*

# Introduction

Quantitative analysis in linguistics has consistently grown in the last few decades, also thanks to the interest of many mathematicians, statisticians and physicists who, using techniques borrowed from statistics and information theory, discovered many structural properties of language streams. Up to now obtained results are still quite modest, but they are still able to discover some previously unknown linguistic features; furthermore new discoveries are always made, in order to contribute to a new and more complete perspective of our understanding on language. In this thesis my main aim is to review some of these results, focusing my analysis on analogies and differences of statistical properties between different languages. We will start from two general linguistic laws, Zipf's and Heaps' laws, and we will later focus on more particular statistical features of texts, burstiness and long-range correlations, whose origins will be studied in this thesis.

In the first chapter, there will be an introduction to two of the most famous laws (not in a rigorous sense, but just from an empirical point of view) in quantitative linguistics: Zipf's law and Heaps' law. Zipf's law, introduced by G. K. Zipf in 1949, studies the relation between the rank of a word $r$ (the position in a classification of all used words, ordered in decreasing order by their frequency) and its frequency $f(r)$ ($f(r) \propto r^{-z}$), while Heaps' law, introduced by H. S. Heaps in 1978, studies the relation between the number of different words $N(k)$ and the total number of used words $k$ ($N(k) \propto k^{\gamma}$). Along this chapter we will study the relation between these two laws, ob-

taining a particular asymptotic relation between the two exponents $z$ and $\gamma$. Moreover we will analyze some interesting models for creating random texts which, using completely different approaches, exhibit these laws. In particular the last model we will study, proposed by M. Gerlach and E. G. Altmann in 2013, is very interesting, in fact this model merges many different ideas present in various previous models, obtaining results that seem consistent with real data.

In the second chapter we will analyze Zipf's and Heaps' laws from an experimental point of view. In the first part we will observe results of how Zipf's and Heaps' laws fit on a series of random texts created using Simon's model, one of the models studied in the first chapter, and using monkey texts. In the second part we will analyze how Zipf's and Heaps' laws fit on real texts, using *War and Peace* in four different languages: English, French, German and Italian. As said before, we will compare these results for all languages studied, observing different and similar behaviors.

In the third chapter there will be a theoretic introduction to a recent model (2012) proposed by E. G. Altmann, G. Cristadoro and M. Degli Esposti, that, building a hierarchy of language, whose levels are established from sets of both semantically and syntactically similar conditions, studies the origins of long-range correlations in texts and how long-range correlations and burstiness behave moving up and down in the hierarchy built. Moreover, after a detailed explanation of this model, there will be a statistical analysis that will be used in the experiments for the approximation of the long-range correlations exponent ($\sigma_X^2(t) \propto t^\gamma$), necessary when working with finite time sequences, as in the case of texts.

This model will be better analyzed thanks to experiments on real texts in the fourth chapter and, even in this case, will be used *War and Peace* in four different languages: English, French, German and Italian. After a pre-

liminary study language per language, in which we will observe long-range correlations and burstiness for each book, there will be a combined analysis in order to observe differences and analogies of long-range correlations and burstiness between different languages.

Finally in the fifth chapter there will be a comparison between two different approaches for quantitative analysis on texts: the one studied along this thesis, and the other one, studied and analyzed by two colleagues of mine, Filippo Bonora and Giulia Tini, in their theses, that consider a text as a network. After an introduction to their method, results obtained for various books with these two different approaches will be shown.

# Contents

# List of Figures

# Chapter 1

# Zipf's and Heaps' laws: theory

In qualitative studies on language the frequency with which different words are used in writing or in speech is clearly the most elementary statistical property of human language. That's why it has been the first to be quantitatively characterized and the most studied. Therefore in this chapter we will analyze two of the most important laws based on frequency: Zipf's and Heaps' laws.

## 1.1 Zipf's law

In this section we will study Zipf's law and a mathematical model whose purpose is to explain it.

### 1.1.1 Origins

In his book *Human Behavior and the Principle of Least Effort*, published in 1949, George Kingsley Zipf, a philologist, proposed a principle, the so colled Principle of Least Effort, that argue that a person will always *"strive to solve his problems in such a way as to minimize the total work that he must expend in solving both his immediate problems and his probable future problems"*. In his explanation, which clearly lacked of a mathematical formulation, he revisited a finding which he had already advanced more that

a decade earlier in his *The Psycho-Biology of Language* (Zipf, 1936), now known as Zipf's law.

Its original formulation establishes that, in a sizable sample of language (a text or a speech) the number of words $K(n)$ which occur exactly $n$ times decays with $n$ as

$$K(n) \propto n^{-\zeta} \tag{1.1}$$

for a wide range of values of $n$. The exponent $\zeta$ changes from text to text but it was often found that $\zeta \sim 2$.

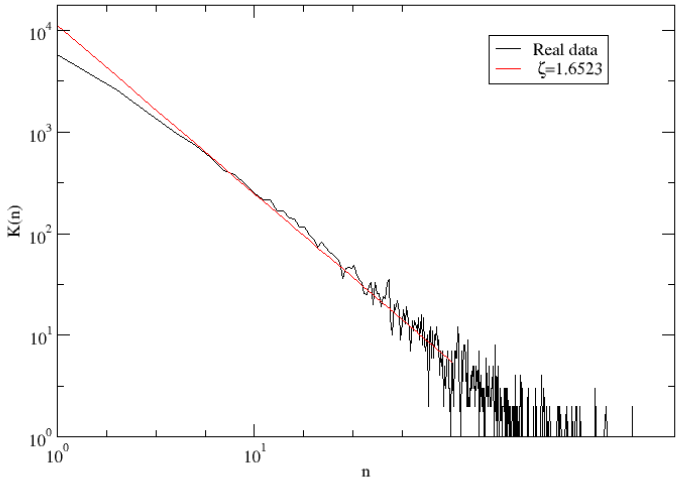



Figure 1.1: $K(n)$ in a log-log plot for the book *War and Peace* in English

Later, in the book *Human Behavior and the Principle of Least Effort*, Zipf proposed an alternative, but equivalent to the first one (we will prove this in a while), formulation: if we rank the words of a chosen text in decreasing order by their frequency (with rank 1 the most frequent word, at rank 2 the second most frequent word and so on), we can observe that the frequency $f$ of the word with rank $r$ follows, to a good approximation, the following

relation with $r$:

$$f(r) \propto r^{-z} \tag{1.2}$$

where it was often found that $z \sim 1$.

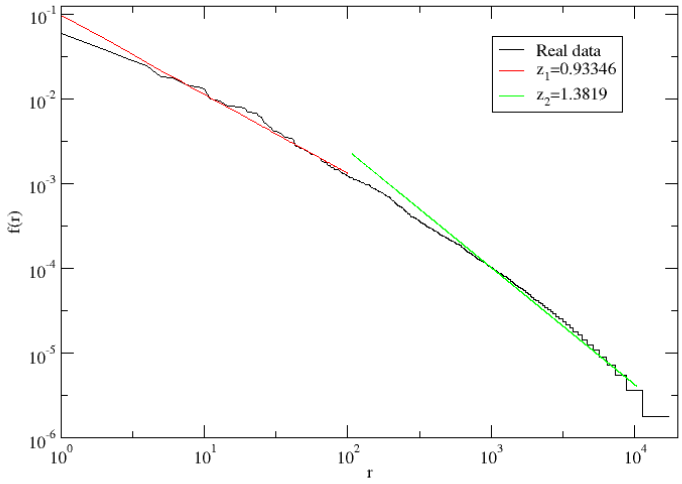



Figure 1.2: $f(r)$ in a log-log plot for the book *War and Peace* in English

**Proposition 1.1.1.** *The equations* $K(n) \propto n^{-\zeta}$ *and* $f(r) \propto r^{-z}$ *are equivalent.*

*Proof.* The rank of a word with $n$ occurrences is equal to the number of words with $n$ or more occurrences: $r(n) = \sum_{n'=n}^{+\infty} N(n') \approx \int_n^{+\infty} N(n')dn'$ where $N(n)$ is the number of words with appears exactly $n$ times.

If $N(n)$ satisfies the first equation, then $f(r)$ satisfies the second one and there is a relation between the two coefficients $\zeta$ and $z$:

$$z = \frac{1}{\zeta - 1}. \tag{1.3}$$

Vice-versa, if $f(r)$ satisfies the second equation, then $N(n)$ satisfies the first one and the following relation is valid:

$$\zeta = 1 + \frac{1}{z} \tag{1.4}$$

□

Zipf proposed a qualitative explanation of this relation between the number of words and the number of occurrences using the Principle of Least Effort, on the basis of the equilibrium between the "work" done by the two agents involved in a communication event: the speaker and the hearer. From the speaker's point of view, the most economic vocabulary consists of a single word that contains all the desired meanings to be verbalized. The hearer, on the other hand, "would be faced by the impossible task of determining the particular meaning to which the single word in a given situation might refer" [2]. This conflict between the speaker's and the hearer's tendencies to respectively reduce and increase lexical diversification, is solved by developing a vocabulary where a few words are used very often, while most words occur just a few times.

### 1.1.2 A mathematical model for Zipf's law

Despite its apparent robustness, Zipf's law is just an empirical observation and not a law in a rigorous sense: in fact this law has been assumed but never explained in models for the evolution of communication. Moreover it can be observed that this law works well for the smallest ranks, but it doesn't fit for bigger ranks (see Fig. 1.2).

In order to solve the second problem, R. F. i Cancho and R. V. Solè [20], observed, analyzing the rank ordering plot of their data (they used BNC, *British National Corpus*, a corpus of modern English, both spoken -10%- and written -90%-), the presence of two different exponents in the same rank ordering plot. The first exponent $z_1 \sim 1$ for ranks $r < b \in \left(10^3, 10^4\right)$ and

the second one $z_2 \sim 2$ for ranks $r > b$. Thus the frequency of words follows the following mathematical law: a double power law, composed by the initial Zipf's law and a more sloping decay.

$$\begin{cases} f(r) \propto r^{-z_1} & r < b, \\ f(r) \propto r^{-z_2} & r > b \end{cases} \tag{1.5}$$

The presence of this double power law can be easily observed in the following figure, caught from [20].

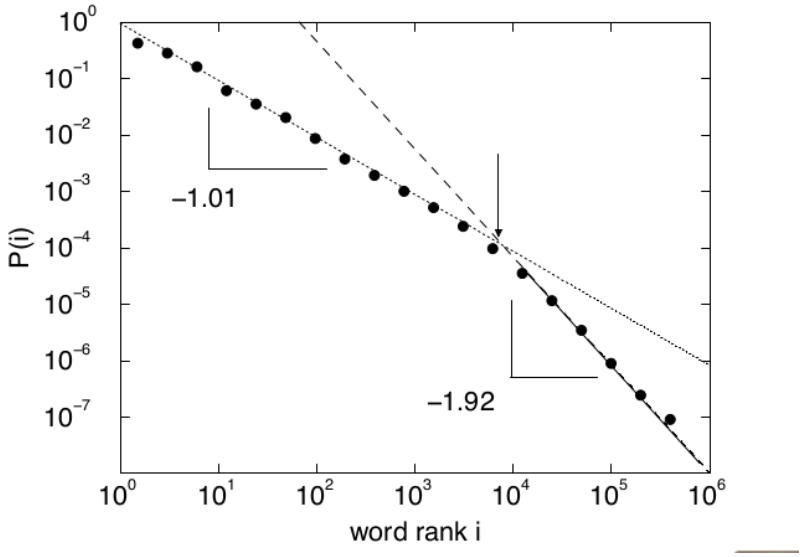



Figure 1.3: Probability of a word as a function of its rank $i$,$P(i)$. The first and the second power law decays have exponent $z_1 = 1.01 \pm 0.02$ and $z_2 = 1.92 \pm 0.07$, respectively. Statistics on the whole BNC, that has a lexicon of 588030 words.

The two observed exponents divide words in two different sets: a kernel lexicon formed by $\approx b$ versatile words and an unlimited lexicon for specific communication. The existence of a kernel lexicon raises the issue of how small can be a lexicon without drastically pauperize communication. Some examples of languages with a very small lexicons are pidgin languages. Pidgin languages are "on-the-spot" languages that develop when people with

no common language come into contact with each other. For example, the establishment of plantation economies in the Caribbean, with large groups of slaves from different language backgrounds, gave life to a number of pidgins based on English, French, Spanish, Dutch, and Portuguese. Estimates of the number of words of the kernel lexicon of a pidgin language vary from about 300 to 1500 words and, as expected, words of such small lexicons are very multi-functional and circumlocutions are often used in order to cover this lexical gap. On the contrary the number of words of the kernel lexicon is about $25000 - 30000$ for ordinary languages, clearly not enough for the 588030 words of BNC. So they suggested that the finiteness of this kernel lexicon is hidden by an unlimited specific lexicon. In fact, although the size of the lexicon of a speaker can be extremely big, what counts for a successful communication are the common words shared with the maximum number of speakers, that is to say, the words in the kernel lexicon.

Now that we have a mathematical law which fits quite good real data, we have to solve the main problem, the complete lack of a quantitative model that explains the process by which a vocabulary diversifies and communication evolves under the pressure of the Principle of Least Effort on both speaker and hearer. In 2003 R. F. i Cancho and R. V. Solè proposed [27] a new explanation mathematical model.

In this model the process of communication implies the exchange of information from a set of $m$ objects of reference, the *meanings*, $R = \{r_1, ..., r_m\}$, using a set of $n$ signals, the *words*, $S = \{s_1, ..., s_n\}$. The interactions between meanings and words (a word can have different meanings and various meanings can be expressed with different words) can be modeled with a binary matrix $A = \{a_{i,j}\}$, where $1 \leq i \leq n$, $1 \leq j \leq m$ and $a_{i,j} = 1$ if the $i$th word refers to the $j$th meaning and $a_{i,j} = 0$ otherwise. We define $p(s_i)$ and $p(r_j)$ as the probability of $s_i$ and $r_j$, respectively and $p(s_i, r_j)$ as the joint probability.

If synonymy is forbidden, we would have

$$p(s_i) = \sum_j a_{i,j} p(r_j), \tag{1.6}$$

because words are used for referring to meanings. We assume $p(r_j) = 1/m$ hereafter. If synonymy is allowed, the frequency of a meaning has to be divided between all its words. The frequency of a word $p(s_i)$ is defined as

$$p(s_i) = \sum_j p(s_i, r_j). \tag{1.7}$$

According to Bayes theorem we have

$$p(s_i, r_j) = p(r_j) p(s_i|r_j) \tag{1.8}$$

and

$$p(s_i|r_j) = a_{i,j} \frac{1}{\omega_j} \tag{1.9}$$

where $\omega_j = \sum_i a_{i,j}$ is the total number of synonyms of the $j$th meaning. Combining the last two equations, we get

$$p(s_i, r_j) = a_{i,j} \frac{p(r_j)}{\omega_i} \tag{1.10}$$

and thus

$$p(s_i) = \sum_j p(s_i, r_j) = \frac{1}{m} \sum_j \frac{a_{i,j}}{\omega_j} \tag{1.11}$$

The effort for the speaker will be defined in terms of the diversity of words, here measured by means of the word entropy

$$H_n(S) = -\sum_{i=1}^{n} p(s_i) \log_n p(s_i). \tag{1.12}$$

So if a single word is used for every meaning, the effort is minimal and $H_n(S) = 0$. Indeed

$$p(s_i) = \begin{cases} 0, & \text{if } i \neq \bar{i} \\ 1, & \text{if } i = \bar{i} \end{cases}$$

and so $p(s_i)\log_n p(s_i) = 0,\ \forall i$, so $H_n(S) = 0$. Vice-versa when all words have the smallest ($\neq 0$) possible frequency ($\frac{1}{n}$), then the frequency effect is in the worst case for all words $\Rightarrow H_n(S) = 1$.

The effort for the hearer when $s_i$ is heard, is defined as

$$H_m(R|s_i) = -\sum_{j=1}^{m} p(r_j|s_i)\log_m p(r_j|s_i) \tag{1.13}$$

where $p(r_j|s_i) = \dfrac{p(r_j, s_i)}{p(s_i)}$ by the Bayes theorem. The effort for the hearer is defined as the average effort for all possible words he can hear, that is

$$H_m(R|S) = -\sum_{i=1}^{n} p(s_i)H_m(R, s_i). \tag{1.14}$$

An energy function combining the effort for both the speaker and the hearer is defined as

$$\Omega(\lambda) = \lambda H_m(R|S) + (1-\lambda)H_n(S), \tag{1.15}$$

where $0 \leq \lambda,\ H_m(R|S),\ H_n(S) \leq 1$. In this way the energy function depends on a single parameter $\lambda$, which represents the contribution of each term to the total effort.

R. F. i Cancho and R. V. Solè performed numerical simulations for $n, m = 150$ where, at each step, a few elements of the matrix $A$ were switched between 0 and 1 or vice-versa, and the change was accepted if the energy function $\Omega(\lambda)$ decreased. They expected that, if Zipf's hypothesis were valid, the probabilities $p(s_i)$ would converge to a distribution compatible with the inverse relation between frequency and rank for some intermediate value of $\lambda$. As a measure of communication accuracy, they also recorded the mutual information between the probability distributions of words and meanings, defined as

$$I(S, R) = \sum_{j=1}^{m} p(r_j)\sum_{i=1}^{n} p(s_i|r_j)\log_n p(s_i|r_j) - \sum_{i=1}^{n} p(s_i)\log_n p(s_i), \tag{1.16}$$

and the relative lexicon size, $L$, defined as the ratio between the number of effectively used words and the total number of available words $n$.

In the following figure, caught from [27], there is a schematic representation of the results of simulations, after a large number of iterations of the dynamical process of switching the elements of $A$. The left panel shows the word-meaning mutual information $I(S,R)$ as a function of $\lambda$. Two different regimes are clearly identified, separated by a sharp transition at $\lambda^* \approx 0.41$. For $\lambda < \lambda^*$, there is practically no informational correlation between words and meanings, which is to say that communication fails. Accordingly, the relative lexicon size $L$ vanishes. Vice-versa for $\lambda > \lambda^*$, both $I(S,R)$ and $L$ attain significant levels, and approach their maximal values for $\lambda \to 1$.

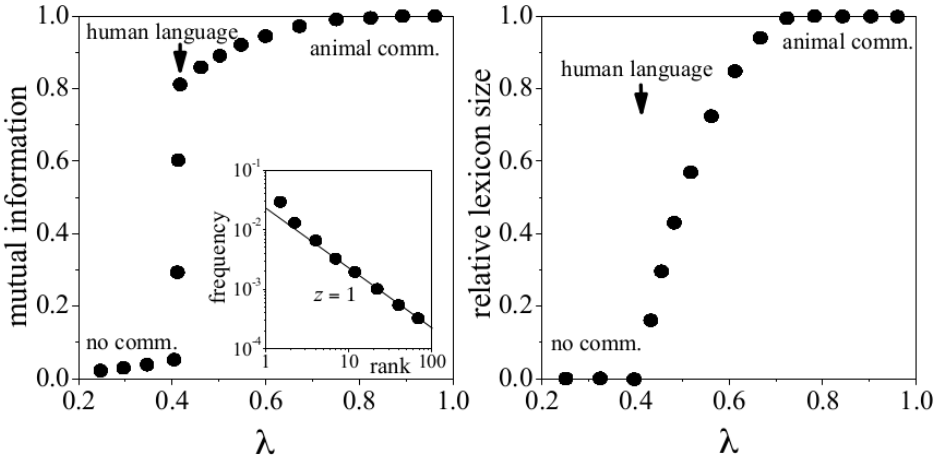



Figure 1.4: Schematic representation of numerical results from the quantitative model for the Principle of Least Effort applied to the evolution of language. Left panel: the word-meaning mutual information $I(S,R)$ as a function of the parameter $\lambda$. Right panel: relative lexicon size $L$ as a function of $\lambda$. Labels indicate the regimes of no communication and animal communication, and the transition at $\lambda^*$, which has been identified with human language.

Moreover, as shown in the insert of the left panel of the figure, the analysis of the frequency-rank relation from the results of simulations satisfy Zipf's law with exponent $z \sim 1$ at the critical value $\lambda^*$, while the power-law relation breaks down for other values of $\lambda$. In the context of this model, human language seems the result of the Principle of Least Effort, that let the system reach the edge of transition.

In conclusion, R. F. i Cancho's and R. V. Solè's evolutionary model demonstrates that a convenient mathematical formulation of the Principle of Least Effort leads to Zipf's law, with $z \sim 1$. However, this result must be interpreted cautiously. In fact, this model describes the evolution of the frequencies of word usage in language as a whole. On the other hand, Zipf's law is known to be valid for single (or a small number of) texts. When many unrelated samples of the same language are joined into a single corpus, the resulting lexicon does not necessarily satisfy Zipf's law, as has been discussed by the same authors in [20].

## 1.2 Heaps' law

In this section we will analyze Heaps' law and two mathematical models that, starting from Zipf's law, show the validity of Heaps' law.

### 1.2.1 Origins

Another important linguistic law is Heaps' law, discovered in the 1960 by Gustav Herdan and later published and better analyzed also by Harold Stanley Heaps. This empirical law describes the number of distinct words, $N$, in a document (or set of documents) as a function of the document length, $k$: the classical result for this relation is the following law:

$$N_k = N(k) \propto k^{\gamma}, \tag{1.17}$$

with $\gamma \in [0, 1]$.

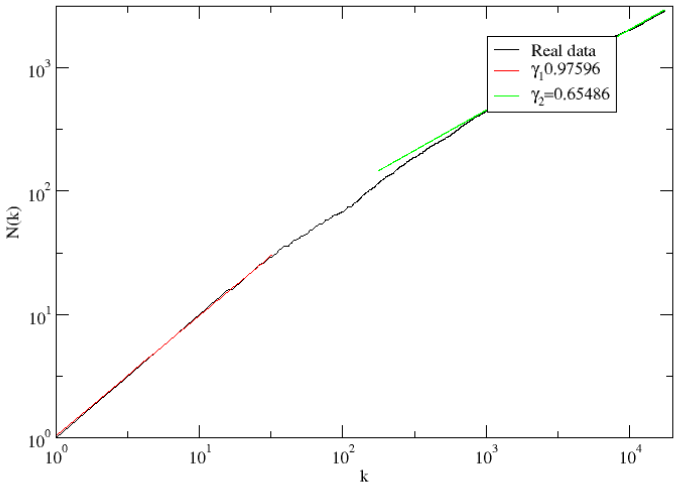



Figure 1.5: $N(k)$ in a log-log plot for the book *War and Peace* in English

Studying this relation, we can note that Heaps' law has the same problems of Zipf's law: there isn't a quantitative model that describes it and real data don't follow a power-law but a double power-law:

$$\begin{cases} N_k \propto k^{\gamma_1} & k \ll M_b, \\ N_k \propto k^{\gamma_2} & k \gg M_b \end{cases} \tag{1.18}$$

where $\gamma_1 \sim 1$, $\gamma_2 \in [0, 1]$, $M_b$ is the number of words such that $N_{M_b} = b$ and $b$ is the same $b$ present in Eq. (1.5) (in the following part of this chapter we will study the relation between Zipf's and Heaps' laws and especially in the last section there will be an explanation of this particulare relation).

The presence of this double power law can be easily observed in the following figure, caught from [45].

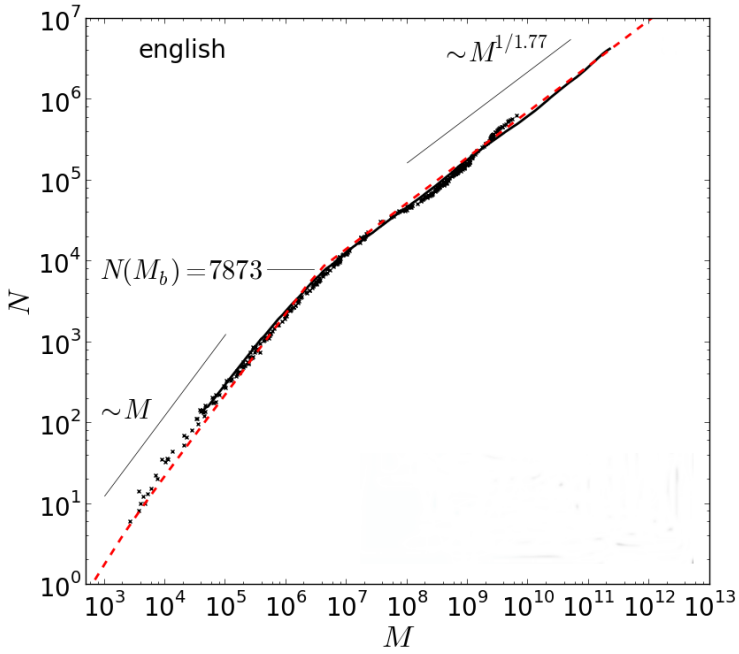



Figure 1.6: Number of different words as a function of the total number of words (in the graph its notation is $M$). The first and the second power law decays have exponent $\gamma_1 = 1$ and $\gamma_2 = \frac{1}{1.77}$, respectively. Statistics on the whole google n-gram database, a corpus of more than over 5.2 million books published in the last centuries and digitized by Google Inc.

### 1.2.2 A formal derivation of Heaps' law

The relation between Zipf's law and Heaps' law is one of the most interesting research argument in linguistic research area. In this part we will study a model with which we will derive Heaps' law directly from the Mandelbrot distribution, which has the original Zipf's law as a special case.

First of all we have to define the Mandelbrot distribution. Given the parameters $N \in \mathbb{Z}$, $c \in [0, +\infty)$ and $\theta \in \mathbb{R}^+$, the Mandelbrot distribution is a discrete probability distribution: given $r \in \{1, 2, ..., N\}$ ($r$ represents the

rank of the data), the probability mass function is given by:

$$f(r; N, c, \theta) = \frac{(r+c)^{-\theta}}{H_{N,c,\theta}}, \tag{1.19}$$

where $H_{N,c,\theta} = \sum_{k=1}^{N} \frac{1}{(k+c)^{\theta}}$.

**Proposition 1.2.1.** *Original Zipf's law is a particular case of the Mandelbrot distribution.*

*Proof.* If we set $c = 0$ and $\theta = 1$ we obtain the following:

$$f(r; N, 0, 1) = \frac{r^{-1}}{H_{N,0,1}} \tag{1.20}$$

which is exactly the original Zipf's law. □

*Notation* 1.2.1. We will often use $p_r = f(r; N, 0, 1) = a_N r^{-1}$, where $a_N = H_{N,0,1}^{-1}$.

Now, starting from these definitions, we can analyze this model, proposed by D.C. van Leijenhorst and Th.P. van der Weide [29] in 2004.

Let $W$ be a set of $N$ words numbered $1, ..., N$ and let $p_i$ the probability that word $i$ is chosen. The underlying text model is the following: words are taken randomly with replacement from the set $W$ according to its probability distribution and we are interested in the asymptotic behavior (for $N \to +\infty$) of the expected resulting number of different words taken. After taking $k$ words $w_1, ..., w_k$ from $W$, let $D_k$ be the set of different words and let $N_k$ be the number of such words, $N_k = \#D_k$. Obviously $N_k \leq N$. We analyze the drawing of the $k$th word for $k > 0$ in detail. There are two possibilities: the $k$th word has been drawn before or not. Let $a \leq k$, then:

$$\begin{aligned} P(N_k = a) &= P\left(N_{k-1} = a - 1 \wedge w_k \notin D_{k-1}\right) + P\left(N_{k-1} = a \wedge w_k \in D_{k-1}\right) = \\ &= P\left(N_{k-1} = a - 1\right) P\left(w_k \notin D_{k-1}\right) + P\left(N_{k-1} = a\right) P\left(w_k \in D_{k-1}\right) \end{aligned} \tag{1.21}$$

Note that $P(N_k = a) = 0,\ \forall a > k$ and $P(N_1 = 1) = 1$.

Finally note that $P(w_k \in D_{k-1}) = 1 - P(w_k \notin D_{k-1})$, where

$$P(w_k \notin D_{k-1}) = \sum_{i \in W} P(w_k = i \wedge i \notin D_{k-1}) = \sum_{i \in W} p_i (1-p_i)^{k-1} \tag{1.22}$$

Hereafter we will use the following notation: $S_k = \sum_{i \in W} p_i (1-p_i)^{k-1}$ and $M_k = \sum_{i \in W} (1-p_i)^k$. We will refer to $M_k$ as the $k$th *reverse moment* of the probability distribution. Then, clearly: $S_k = M_{k-1} - M_k$. Now, if we use the following notation, $N(k,a) = P(N_k = a)$, we get the following recurrence relation:

$$\begin{cases} N(1,1) = 1 & \\ N(k,a) = 0 & \text{if } a > k \\ N(k,a) = N(k-1,a-1)S_k + N(k-1,a)(1-S_k) & \text{if } a \leq k \end{cases} \tag{1.23}$$

Now we can study the expected number of different words $N_k$ after taking $k$ words randomly from the set $W$ of words. Before studying this, we are going to observe the following lemma.

**Lemma 1.2.2.** *The expected number of different words in a radom selection of $k$ words is $N_k = N - M_k$.*

*Proof.* From the previous recurrence relation,

$$N_k = \sum_{a=1}^{k} aN(k,a) = \sum_{a=1}^{k} aN(k-1,a-1)S_k + \sum_{a=1}^{k} aN(k-1,a)(1-S_k) = \tag{1.24}$$

$$= S_k \underbrace{\left(\sum_{a=1}^{k} aN(k-1,a-1) - \sum_{a=1}^{k} aN(k-1,a)\right)}_{\sum_{a=1}^{k-1}(a+1)N(k-1,a) - \sum_{a=1}^{k-1} aN(k-1,a)} + \underbrace{\sum_{a=1}^{k} aN(k-1,a)}_{N_{k-1}} = \tag{1.25}$$

$$= S_k + N_{k-1} \tag{1.26}$$

Now we have to show that $N_k = N - M_k$, so it' sufficient to prove that $N_k + M_k = N$. In order to prove this statement we can use the principle of induction.

- $k = 1$: $N_1 + M_1 = 1 + \sum_{i \in W}(1 - p_i) = 1 + N - 1 = N$ ✓
- $k' = k + 1$: $N_{k'} = N_{k'-1} + S_{k'} = N_k + S_{k+1} = N - M_k + S_{k+1} = N - M_k + M_k - M_{k+1} = N - M_{k'}$ ✓.

□

Now, before proving the following theorem, we have to add two hypothesis: $c > 0$ and $1 < \theta \leq 2$.

**Theorem 1.2.3.** *The expected number $N_k$ of different words in a random selection of $k$ words from $N$ is*

$$N_k = \alpha k^{\beta}(1 + o(1)) + O\left(\frac{k}{N^{\theta-1}}\right), \ \left(N, k \to \infty, \ \frac{k}{N^{\theta-1}} \to 0\right) \tag{1.27}$$

*where $\beta = \theta^{-1}$ and $\alpha = a_{\infty}^{\beta}\Gamma(1 - \beta)$ with $a_{\infty} = \lim_{N\to\infty} a_N$ and $\Gamma$ is the well known gamma function, $\Gamma(x) = \int_0^{\infty} t^{x-1}\exp(-t)dt$.*

*Proof.* First of all it's convenient to have these further notations:

$$A = \beta a_N^{\beta}, \ \mu = -1 - \beta, \ t(x) = a_n(c + x)^{-\theta}, \tag{1.28}$$

$$\phi_k(x) = (1 - t(x))^k, \psi(t) = (1 - t)^{k-1}t^{\mu+1}. \tag{1.29}$$

In order to prove this theorem we may split it in points.

- Fixed $k > 0$ and $N$, $\phi_k(x)$ is a monotonically increasing function: $[1, \infty) \to (0, 1)$, so

$$\sum_{i=1}^{N-1} \phi_k(i) \leq \int_1^N \phi_k(x)dx \leq \sum_{i=2}^{N} \phi_k(i). \tag{1.30}$$

  Hence,

$$M_k = \sum_{i=1}^{N} \phi_k(i) = \int_1^N \phi_k(x)dx + \varepsilon \tag{1.31}$$

  with error $|\varepsilon| \leq \phi_k(1) + \phi_k(N)$. Now, because of $a_N$ is uniformly bounded in $N$ and $k$, then even $\varepsilon$ is uniformly bounded in $N$ and $k$. In this way, the reverse moment $M_k$ is approximated by an integral.

- By substitution of $t(x) = a_N(c+x)^{-\theta}$, we have $dx = -At^{\mu}dt$ and

$$\int_1^N \phi_k(x)dx = A\int_{t(N}^{t(1)} (1-t)^k t^{\mu}dt. \tag{1.32}$$

- Integrating by parts the previous integral, we obtain:

$$A\int_{t(N}^{t(1)} (1-t)^k t^{\mu}dt = \left[A\frac{(1-t)^k t^{\mu+1}}{\mu+1}\right]_{t(N)}^{t(1)} - A\theta k\int_{t(N)}^{t(1)} \psi(t)dt. \tag{1.33}$$

Now, the first part of $A\int_{t(N}^{t(1)} (1-t)^k t^{\mu}dt$ is equal to $(c+N)\phi_k(N) - (c+1)\phi_k(1)$. By Taylor expansion $\phi_k(N) = \left(1 - a_N(c+N)^{-\theta}\right)^k = 1 + O\left(\frac{k}{(c+N)^{\theta}}\right)$, $\left(\frac{k}{(c+N)^{\theta}} \to 0\right)$. In this way the first part of $A\int_{t(N}^{t(1)} (1-t)^k t^{\mu}dt$ has been estimated as $O(1) + N + O(\frac{k}{N^{\theta-1}})$. Thus we obtain approximately the number of words $N$ in the set $W$.

The second part of $A\int_{t(N}^{t(1)} (1-t)^k t^{\mu}dt$ is $-A\theta k\int_{t(N)}^{t(1)} \psi(t)dt$ that can be split into three terms as: $-A\theta k\int_0^1 \psi(t)dt + A\theta k\int_0^{t(N)} \psi(t)dt + A\theta k\int_{t(1)}^1 \psi(t)dt$.

1. The second term $A\theta k\int_0^{t(N)} \psi(t)dt = A\theta k\int_0^{t(N)} (1-t)^{k-1}t^{\mu+1}dt$. Since $t(N) \to 0$ if $N \to \infty$, this term has order

$$A\theta k O\left(\int_0^{t(N)} 1\ t^{\mu+1}dt\right) = O(kt(N^{\mu+2})) = O\left(\frac{k}{N^{\theta-1}}\right).$$

2. By partial integration the third term

$$A\theta k\int_{t(1)}^1 \psi(t)dt = \tag{1.34}$$

$$= \left[A\theta k(1-t)^{k-1}\frac{t^{\mu+2}}{\mu+2}\right]_{t(1)}^1 + \frac{A\theta k(k-1)}{\mu+2}\int_{t(1)}^1 (1-t)^{k-2}t^{\mu+2}dt$$

Now, $A$ is bounded and $\mu+2>0$, so $\left[A\theta k(1-t)^{k-1}\frac{t^{\mu+2}}{\mu+2}\right]_{t(1)}^{1} = O(k(1-t(1))^{k-1}),\ (k\to\infty)$. Similarly, $\frac{A\theta k(k-1)}{\mu+2}\int_{t(1)}^{1}(1-t)^{k-2}t^{\mu+2}dt \leq \frac{A\theta k(k-1)}{\mu+2}\int_{t(1)}^{1}(1-t)^{k-2}dt = O((k^2(1-t(1)))^{k-1}),\ (k\to\infty)$. Thus we can observe that the third term decreases exponentially with $k$.

Summarizing: up to now we have proved that

$$M_k = -A\theta k\int_0^1 \psi(t)dt + N + O(1) + O\left(\frac{k}{k^{\theta-1}}\right) \tag{1.35}$$

where $N, k\to\infty,\ \frac{k}{N^{\theta-1}}\to 0$.

- The integral in the previous equation can be recognised as

$$-A\theta k\frac{\Gamma(k)\Gamma(\mu+2)}{\Gamma(k+\mu+2)}, \tag{1.36}$$

valid only for $\mu+2\neq 0 \Leftrightarrow \theta\neq 1$. Now, using Stirling's approximation of the $\Gamma$ function $\left(\Gamma(x+1)\sim\sqrt{2\pi x}x^x\exp(-x)\right)$, we have

$$-A\theta k\frac{\Gamma(k)\Gamma(\mu+2)}{\Gamma(k+\mu+2)}\sim \tag{1.37}$$

$$\sim -A\theta\Gamma(\mu+2)k^{-\mu+1} = -a_N^{\beta}\Gamma(1-\beta)k^{\beta} \tag{1.38}$$

Finally, $a_N^{\beta} = a_{\infty}^{\beta}(1+o(1))$, so we have

$$-A\theta k\int_0^1\psi(t)dt = a_{\infty}^{\beta}\Gamma(1-\beta)k^{\beta}(1+o(1)) \tag{1.39}$$

Substituting this into $M_k = -A\theta k\int_0^1\psi(t)dt + N + O(1) + O\left(\frac{k}{k^{\theta-1}}\right)$ and using $N_k = N - M_k$ we complete the proof of the theorem.

□

An immediate consequence of this theorem is Heaps' law.

**Corollary 1.2.4** (Heaps' law)**.** $N_k = \alpha k^\beta$ *for* $k, N \to \infty$.

One of the main results of this model is the relation between Zipf's coefficient $z$ and Heaps' coefficient $\gamma$. In fact in this model $z$ and $\gamma$ are relatively $1 < \theta \leq 2$ and $\beta$, where $\beta = \theta^{-1}$.

### 1.2.3 From Zipf's law to Heaps' law

In the last subsection we have analyzed a stochastic model which, starting from Zipf's law, leads to Heaps' law. Now we will analyze a recent discovery, published in 2010 by L. Lü, Z.-K. Zhang, T. Zhou [39] and our goal will be to prove that for an evolving system with a stable Zipf's exponent, Heaps' law can be directly derived from Zipf's law without the help of any specific stochastic model. Morover the relation $\gamma = \dfrac{1}{z}$ is only an asymptotic solution hold for very large size systems with $z > 1$. This model also refines this result for finite size systems with $z \gtrsim 1$ and complete it with $z < 1$.

First of all we can note that $r - 1$ is the number of distinct words with frequency larger than $f(r)$. So, denoting by $k$ the total number of word occurrences and $N_k$ the corresponding numbers of distinct words

$$r - 1 = \int_{t(r)}^{t(1)} N_k p(t')dt' \tag{1.40}$$

where $t(r) = f(r)k$ is the number of occurrences of a word of rank $r$.

Remembering from Eq. (1.1) that $p(t) = At^{-\zeta}$ with $A$ constant and according to the normalization condition $\displaystyle\int_1^{t(1)} p(t)dt = 1$,

$$A = \frac{\zeta - 1}{1 - t(1)^{1-\zeta}} \approx (\text{ when } \zeta > 1 \text{ and } t(1) \gg 1) \approx \zeta - 1. \tag{1.41}$$

Substituting $p(t')$ in the equation before by $(\zeta - 1)t'^{-\zeta}$, we have

$$r - 1 = N_k \left[t(r)^{1-\zeta} - t(1)^{1-\zeta}\right]. \tag{1.42}$$

According to Zipf's law and the relation between the Zipf's and power-law exponents $\zeta = 1 + \frac{1}{z}$, the right part of the last equation can be written in the following way

$$r - 1 = N_k \left[ t(1)^{-\frac{1}{z}} (r-1) \right]. \tag{1.43}$$

Combining $r - 1 = \int_{t(r)}^{t(1)} N_k p(t') dt'$ and $r - 1 = N_k \left[ t(1)^{-\frac{1}{z}} (r-1) \right]$ we can obtain the estimation of $t(1)$, as

$$t(1) = N_k^z. \tag{1.44}$$

Obviously, the text size $k$ is the sum of all words' occurrences, that is to say

$$k = \sum_{r=1}^{N_k} t(r) \approx \int_1^{N_k} t(r) dr = \frac{t(1)\left(N_k^{1-\alpha} - 1\right)}{1-\alpha} \tag{1.45}$$

Note that the summation $\sum_{r=1}^{N_k} t(r)$ is larger than the integration $\int_{r=1}^{N_k} t(r) dr$ but $\sum_{r=1}^{N_k} t(r)$ can be approximated with $\int_{r=1}^{N_k} t(r) dr$ because the relative error of this approximation $\frac{\sum_{r=1}^{N_k} t(r) - \int_{r=1}^{N_k} t(r) dr}{\sum_{r=1}^{N_k} t(r)}$ increases with the increasing of $z$ and decreases with the increasing of $N$.

Now it's clear the following relation:

$$k = \frac{N_k \left(N_k^{1-z} - 1\right)}{1-z}. \tag{1.46}$$

This equation is clearly not a simple power law form as described in Heaps' law, but we will see that Heaps' law is an approximate result that can be derived from this. Actually, when $z$ is considerably larger than 1, $N_k^{1-z} \ll 1$ and $N_k \approx [k(z-1)]^{\frac{1}{z}}$, while if $z$ is considerably smaller than 1, $N_k^{1-z} \approx (1-\alpha)k$. This approximated result can be summarized as

$$\gamma = \begin{cases} \frac{1}{z}, & z > 1, \\ 1, & z < 1, \end{cases} \tag{1.47}$$

Although Eq. (1.46) is different from a power law, numerical results indicate that the relationship between $N_k$ and $k$ can be well fitted by a power law function.

To validate these numerical results we can propose a stochastic model. Given the total number of occurrences $k$, there are at most $k$ distinct words that may appear. The initial occurrence number of each of these $k$ words is set at 0. At each time step, these $k$ words are sorted in decreasing order by their frequency and the probability a word with rank $r$ will occur in this time step is proportional to $r^{-z}$. The whole process stops after $k$ time steps. The distribution of word occurrence clearly follows Zipf's law with a stable exponent $z$, and the growth of $N_k$ approximately follows the Heaps' law with $\gamma$ dependent on $z$. In the following figure, caught from [39] it's clearly shown how the simulation results of this stochastic model about $\gamma$ vs. $z$ strongly support the validity of Eq. (1.46).

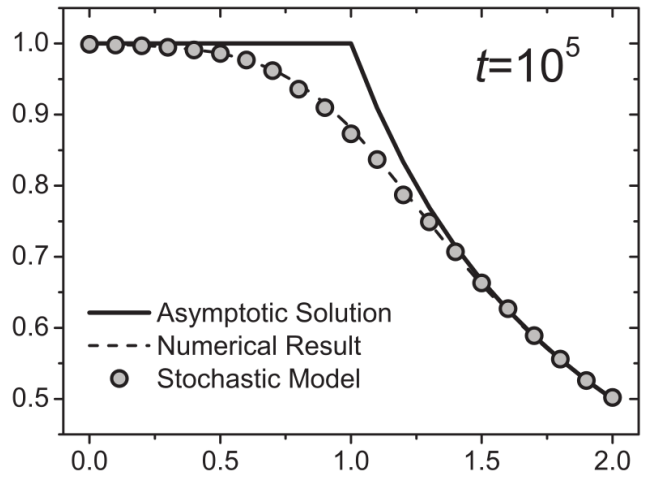



Figure 1.7: Relationship between the Zipf's exponent $z$, $x$ axis, and the Heaps' one $\gamma$, $y$ axis. For the numerical result and the result of the stochastic model, the total number of word occurrences is fixed at $k = 10^5$.

## 1.3 Statistical model for vocabulary growth

In this section we will analyze a new model, proposed in 2013 by M. Gerlach and E. G. Altmann [45], for the number of different words in a given database. The main feature of this model is the existence of two different classes of words: a finite number of core-words which have higher frequency and don't influence the probability of a new word to be used, and the remaining potentially infinite number of noncore-words which have lower frequency and, once used, reduce the probability of the appearance of a new word. This model is based on an analysis of the google-n-gram database (a corpus of more than over 5.2 million books published in the last centuries and digitized by Google Inc) and its main result is the generalization of Zipf's and Heaps' laws to double power law regimes, Eq.(1.5) and Eq. (1.18).

But before studying specifically this model, it's useful to introduce Simon's model, proposed in [3] and later better analyzed in [42].

### 1.3.1 Simon's model

In 1955 the sociologist H. A. Simon [3] proposed a model, based on a multiplicative stochastic process for the recurrent use of words and its main feature is that it is able to quantitatively exhibit Zipf's law.

Here it follows the rigorous explanation of this model.
Consider the process of text generation as a sequence of events where one word is added at each step, and let $K_k(n)$ be the number of different words that appear exactly $n$ times when the text has reached a length of $k$ words. Simon's model proposes the following two dynamical rules for each step:

- with constant probability $\alpha$, the word added at step $k+1$ is a new one, which has not occured in the first $k$ steps. Namely, new words appear in a text at a constant rate $\alpha$;

- the probability that the $(k+1)$th word that has already appeared exactly $n$ times is proportional to $nN_k(n)$, that is to the total number of occurrences of all the words that have appreared exactly $n$ times.

This latter rule can be modified and substituted with the following rule, alternative and equivalent to the previous:

- with probability $1-\alpha$ the word added at step $k+1$ is one of the words that have already occurred in the text; this recurrent word is chosen with a probability proportional to the number of its previous appearances.

Approximating the expectation value of the number of different words with exactly $n$ at step $k+1$ by $K_{k+1}(n)$ itself, these rules let us write the following recursive equation for $K_k(n)$:

$$\begin{cases} K_{k+1}(1) - K_k(1) = \alpha - \frac{1-\alpha}{k} K_k(1), & \text{for } n = 1 \\ K_{k+1}(n) - K_k(n) = \frac{1-\alpha}{k} \left[ (n-1)K_k(n-1) - nK_k(n) \right], & \forall n > 1 \end{cases} \tag{1.48}$$

The first term in the right part of the first equation represents the contribution to $K_{k+1}(1)$ of the word that appears for the first time at step $k+1$. Other terms in both equations are gain and loss contributions associated to the appearance of a word with, respectively, $n-1$ and $n$ previous occurrences. Thanks to this formulation, Simon's model can be seen as a dynamical system for the function $K_k(n)$, where $k$ plays the role of a discrete "time" variable. The above equations for $K_k(n)$ should be solved for a given "initial condition", $N_{k_0}(n)$, which represents the distribution of occurrences of the words that have already been addes to the text at the point $k_0$ at which the model's dynamical rules begin to act.

Equations in (1.48) don't have a stationary solution, in the sense that an asymptotic $k$-indipendent form for $K_k(n)$ doesn't exist. In fact, as $k$ grows, then obviously $k = \sum_n nK_k(n)$ must increase accordingly. A "steady-state"

solution, however, can be reached by assuming that for large $k$, the number of different words with $n$ occurrences satisfies

$$\frac{K_{k+1}(n)}{K_k(n)} = \frac{k+1}{k}, \ \forall n. \tag{1.49}$$

This is equivalent to postulate the existance of a stationary profile $P(n)$ for $K_k(n)$ such that $K_k(n) = kP(n)$. Indeed equations in Eq. (1.48) produce $k$-independent equation for $P(n)$, whose solution is

$$P(n) = \frac{\alpha}{1-\alpha}\beta(n, \zeta), \tag{1.50}$$

where $\beta$ is the Beta function and $\zeta = 1 + (1-\alpha)^{-1}$.

For small values of $\alpha$ ($\lesssim 0.1$) and $\forall n \geq 1$, the above solution for the profile $P(n)$ is very well approximated by the power-law function

$$P(n) \approx \frac{\alpha}{1-\alpha}\Gamma(\zeta)n^{-\zeta}, \tag{1.51}$$

where $\Gamma(\zeta)$ is the Gamma function. This leads for $K_k(n)$ the form given by Zipf's law, Eqs. (1.1) and (1.2) with $z = 1 - \alpha$. Since the probability of appearance of new words must necessarily be larger than 0, the exponent of the frequency-rank relation predicted by Simon's model is always: $z < 1$. The characteristic value $z = 1$ is obtained in the limit $\alpha \to 0$, when the appearance of new words becomes extremely rare, condition expected as the text grows and becomes longer and longer.

Note that $K_k(n) = kP(n)$, with the profile $P(n)$ given by Eq. (1.50) is an exact solution to Simon's model equations: it doesn't represent a general solution, but a solution just for a specific initial condition $K_{k_0}(n) = k_0 P(n)$ which already exhibits the profile $P(n)$. Due to the linearity of Eqs. (1.48), the general solution to Simon's model is a sum of the above special solution, $kP(n)$, plus a contribution from the initial condition.

Hence, Simon's model predicts that a power-law dependance between number of words, occurrences and ranks should hold for small to moderately

large values of $n$, or, in other words, for the lower ranks in Zipf's word list (large $r$). For the higher ranks, on the other hand, deviations from Zipf's law are expected.

### 1.3.2 Gerlach's and Altmann's model

Now that we have introduced Simon's model, we can analyze Gerlach's and Altmann's model which generalizes Simon's one. Its main innovation is that it uses in a statistical model the idea of the presence of two distinct classes of words, idea already present (but studied only from a qualitative point of view) in the work, cited before, of R. F. i Cancho and R. V. Solè [20].

Here it follows the rigorous explanation of this model.
At each step a word is drawn ($k \to k+1$) and the choice of the word follows the rules specified below. The total number of different words is given by $N = N_c + N_{\bar{c}}$, where ($N_{\bar{c}}$) $N_c$ is the number of (non)core-words. The drawn word can either be a new word ($N \to N+1$) with a probability $p_{new}$ or an already existing word ($N \to N$) with probability $1 - p_{new}$. In the latter case, a previously used word is chosen with probability proportional to the number of times this word has occurred before. In the former case, the new word can either originate from a finite set of $N_c^{max}$ core-words ($N_c \to N_c + 1$) with probability $p_c$ or come from a potentially infinite set of noncore-words ($N_{\bar{c}} \to N_{\bar{c}} + 1$).

In the simplest model, we consider $p_c$ to be a constant, that is $p_c^0 \lessapprox 1$, which becomes 0 only if all core-words are drawn ($N_c = N_c^{max}$):

$$p_c(N_c) = \begin{cases} p_c^0, & \text{if } N_c < N_c^{max} \\ 0, & \text{if } N_c = N_c^{max} \end{cases} \tag{1.52}$$

The final element of this model, which establishes the distinguishing aspects of core-words, is the dependence of $p_{new}$ on $N$. So we choose $p_{new}$ depending from $N$ and not from $k$ because an increase in $N$ necessarily implies that fewer undiscovered words exist, while an increase in $k$ is strongly

affected by repetitions of frequently used words.

By definition, we think of core-words as necessary in the creation of any text and, therefore, the use of a new core-word in a particular text should be expected and thus not affect the probability of using a new noncore-word in the future, that is $p_{new} = p_{new}(N_{\overline{c}})$. On the other hand, if a noncore-word is used for the first time ($N_{\overline{c}} \to N_{\overline{c}} + 1$) the combination of this word with the previously used (core and noncore) words lead to a combinatorial increase in possibilities of expression of new ideas with the already used vocabulary and thus to a decrease in the marginal need for additional new words. This argument hints that $p_{new}$ should decrease with $N_{\overline{c}}$.

Considering these factors, we can propose an update rule for $p_{new}$ after each occurrence of a new noncore word as

$$p_{new} \to p_{new} \left( 1 - \frac{\alpha}{N_{\overline{c}} + s} \right), \tag{1.53}$$

with the decay rate $\alpha$ and the constant $s \gg 1$ which is introduced simply in order to muffle the decrease of $p_{new}$ for small $N_{\overline{c}}$ (for simplicity, we use $s = N_c^{max}$). The main justification for the exact functional form in Eq. (1.53) is that it allows us to recover the empirical observations.

In the following figure there is a brief explanation of this model.

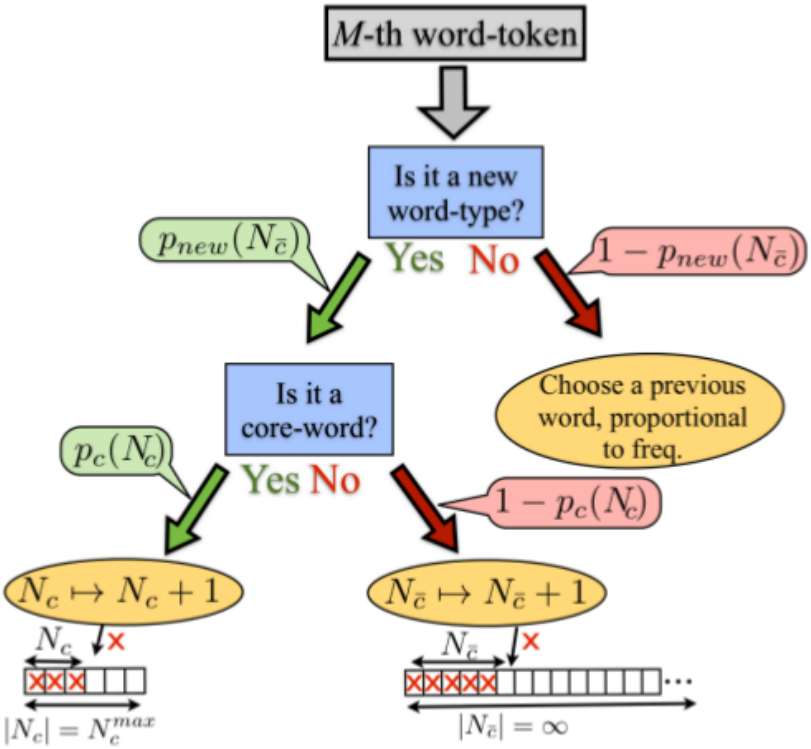



Figure 1.8: Illustration of this generative model forthe usage of new words ($M = k$)

Now we can show how this model implies the validity of Eq. (1.5) and (1.18). We require that $1 - p_c^0 \ll 1$, which simply means that it is much more likely to draw core-words than noncore-words initially. In this case we can obtain approximately exact solutions for $N_k$ in the two limiting cases considered in Eq. (1.18). When $N \ll N_c^{max}$, which implies $N_c, N_{\bar{c}} \ll N_c^{max}$, it follows from Eq. (1.52) and (1.53) that $p_{new} \approx c$, with $c$ constant, so we obviously get: $N \propto k^1$. This case describes the very beginning of the vocabulary growth, when most of new words belong to the set of core-words. In the case $N \gg N_c^{max}$, $p_c = 0$ and $N \approx N_{\bar{c}}$, Eq. (1.53) becomes in the

continuum limit:

$$\frac{d}{dN}p_{new}(N) = -\alpha\frac{p_{new}(N)}{N} \tag{1.54}$$

for which it follows that $p_{new} \propto N^{-\alpha}$.

We now obtain the expected growth curve $N_k$. Note that this model can be considered a biased random walk in $N$, which, as an approximation, can be mapped into a binomial random walk by the coordinate transformation $N_k$ such that $p_{new}(N) = p_{new}(N_k)$. The resulting Poisson-binomial process can be treated analytically, so $N_k$ can be given by the average of the vocabulary growth:

$$N_k = \int_0^k p_{new}(k')dk' = \int_{N_0}^{N_k} p_{new}(N')\left|\frac{dk'}{dN'}\right| dN'. \tag{1.55}$$

Using $p_{new} \propto N^{-\alpha}$, this equation holds by assuming a sub-linear growth for the vocabulary $N \propto k^{\lambda}$, where the relation $\lambda = (1+\alpha)^{-1}$ is established. Now we can identify the following relation between the parameters: $N_c^{max} = b$ and $\alpha = z - 1$. The fitting parameters of Eq. (1.5) can thus be interpreted as: $b$ is the size of the core vocabulary and $z$ controls the sensitivity of the probability of using new words to the number of already used words in Eq. (1.54).

# Chapter 2

# Zipf's and Heaps' law: experiments

Up to know we have studied research evolution and some mathematical models about Zipf's ad Heaps' law. Now we will focus on some experiments on these two laws. First of all we will study two models for creating random texts that exhibit Zipf's and Heaps' laws, while, in the last part of the chapter we will analyze these two laws on real texts (we will use *War and Peace* by Leo Tolstoj), focusing our study on eventual differences between different languages.

## 2.1 Random texts

In this section we will analyze two different kind of random texts, the first one created using Simon's model, already explained in the previous chapter, and the second one created using monkey texts, focusing our study on Zipf's and Heaps' laws.

### 2.1.1 Experiments on Simon's model

In subsection **1.3.1** we analyzed a model for the creation of a random text which exhibits Zipf's law and we will see directly, thanks to some experimental results, what we have just proved in the previous subsection.

I wrote a program which, using different values of $\alpha$, creates random texts using conditions of Simon's model and corresponding Zipf's and Heaps' plots. In fact Simon's model exhibits even Heaps' law, as we can see in the following proposition.

**Proposition 2.1.1.** *Simon's model exhibits Heaps' law with the coefficient of* Eq. (1.16) $\gamma \approx 1$.

*Proof.* We know that the probability of a new word is $\alpha$, so, using the notation $\mathbb{E}(X)$ as the expected value of the random variable $X$, we have that $\mathbb{E}(N_k) = \mathbb{E}(N_{k-1}) + \alpha$ and, given $N(1) = 1$, we can get $\mathbb{E}(N_k) = 1 + \alpha(k-1) = (1-\alpha) + \alpha k$, so $N_k \propto k^1$. □

In the following part of this subsection these results are shown and explained, accordingly to what we have up to now studied.

First of all, we can observe, in the following figures, the behavior of Zipf's and Heaps' laws for two cases studied in my analysis.

*Remark* 2.1.1. For figures and approximation I've used the program *Grace*.

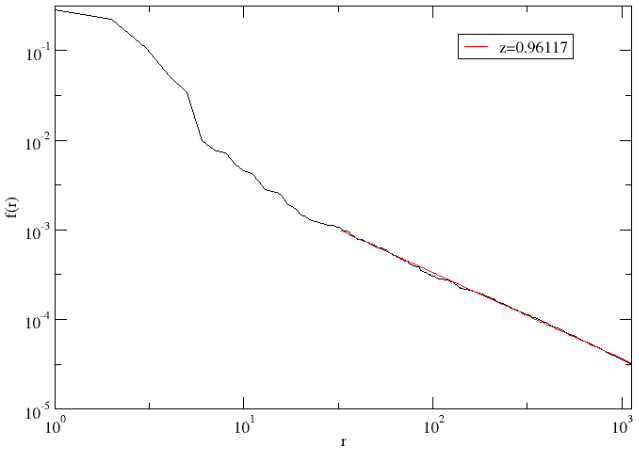



Figure 2.1: Zipf's law, in a $\log - \log$ plot, for a random text of 500000 words, created using Simon's model with $\alpha = 0.04$. $z = 0.96117 \pm 1.0074 \times 10^{-3}$

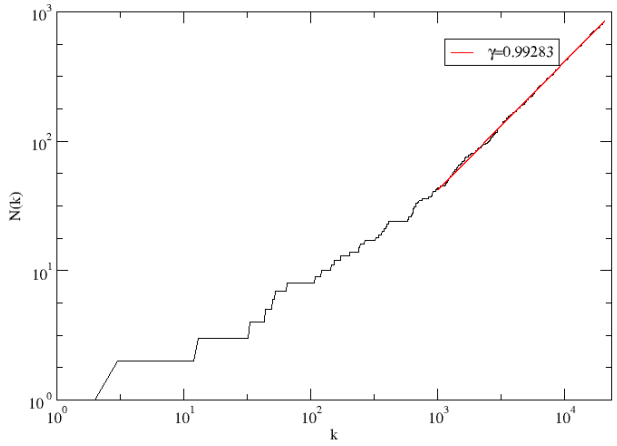



Figure 2.2: Heaps' law, in a $\log - \log$ plot, for a random text of 500000 words, created using Simon's model with $\alpha = 0.04$. $\gamma = 0.99283 \pm 0.15057 \times 10^{-3}$

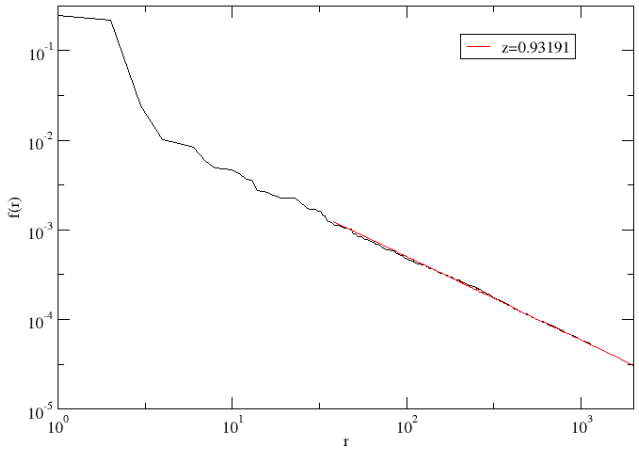



Figure 2.3: Zipf's law, in a $\log - \log$ plot, for a random text of 500000 words, created using Simon's model with $\alpha = 0.08$. $z = 0.93191 \pm 0.41580 \times 10^{-3}$

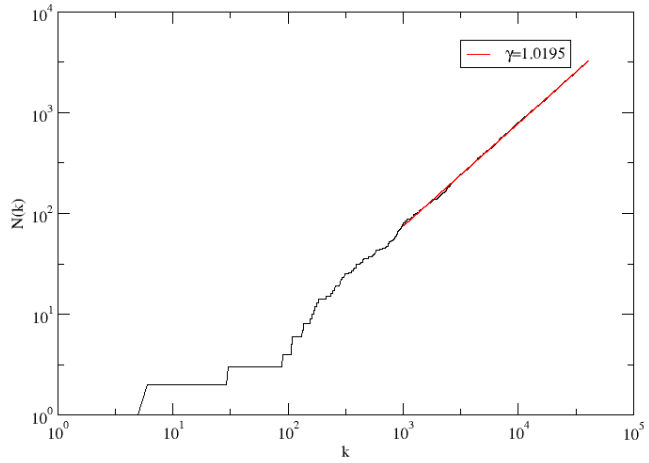



Figure 2.4: Heaps' law, in a $\log - \log$ plot, for a random text of 500000 words, created using Simon's model with $\alpha = 0.08$. $\gamma = 1.0195 \pm 0.72510 \times 10^{-6}$

As predicted in the previous subsection, we can easily note that Zipf's and Heaps' laws are valid and their coefficients follow all the laws we have proved. In fact $z < 1$ for both our cases and $\gamma \approx 1$.

Morover these behaviors confirm the relation between $z$ and $\gamma$, exposed in Eq. (1.46): $z < 1 \Rightarrow \gamma = 1$. These considerations can be observed even better in the following tab, where all results of my experiments on Simon's model are shown.

| $\alpha$ | $z$ | $\gamma$ |
|---|---|---|
| 0.02 | 1.0337 | 1.0504 |
| 0.02 | 0.96564 | 1.0474 |
| 0.02 | 1.0834 | 1.1125 |
| 0.02 | 1.0732 | 0.92077 |
| 0.02 (average) | 1.038985 | 1.0327675 |
| 0.04 | 0.93458 | 1.044 |
| 0.04 | 0.98318 | 1.0153 |
| 0.04 | 0.96117 | 0.99283 |
| 0.04 | 0.95308 | 1.0687 |
| 0.04 (average) | 0.9580025 | 1.0302075 |
| 0.06 | 0.91211 | 1.0694 |
| 0.06 | 0.95761 | 0.98874 |
| 0.06 | 0.97772 | 0.97153 |
| 0.06 | 0.95987 | 1.0004 |
| 0.06 (average) | 0.9518275 | 1.0075175 |
| 0.08 | 0.95182 | 0.98488 |
| 0.08 | 0.94473 | 0.98184 |
| 0.08 | 0.93191 | 1.095 |
| 0.08 | 0.98915 | 0.98634 |
| 0.08 (average) | 0.9544025 | 0.99314 |

The only part in contrast with our study is the case $\alpha = 0.02$, in which Zipf's coefficient $z \not< 1$, but this is probably due to statistical fluctuations for the choice of a small value of $\alpha$. In fact, as we said before, the limit $z = 1$ is reached for $\alpha \to 0$ and the obtained $\gamma \approx 1$.

Now, for concluding our study on Simon's model, we can observe the following figures, where Zipf's and Heaps' plots are shown for all random texts created (4 texts for each $\alpha$).

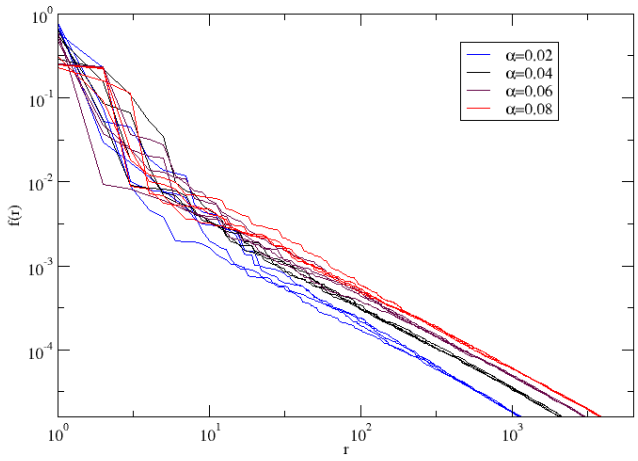



Figure 2.5: Zipf's law, in a $\log - \log$ plot, for all random texts of 500000 words, created using Simon's model

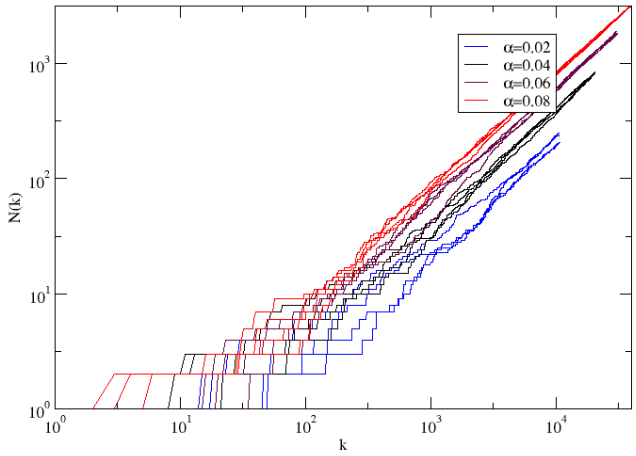



Figure 2.6: Heaps' law, in a $\log - \log$ plot, for all random texts of 500000 words, created using Simon's model

### 2.1.2 Monkey texts

In the last subsection we have observed the results of a model created for exhibiting Zipf's law. Now, on the contrary, we will prove, using [41], that a more general random text, in particular a monkey book, exhibits Zipf's law.

Imagine an alphabet with $A$ letters and a typewriter with a keyboard with one key for each letter and a space bar. For a monkey randomly typing on the typewriter the chance for hitting the space bar is assumed to be $q_s$ and the chance for hitting any of the letters is $\frac{1-q_s}{A}$. A word is then defined as a sequence of letters surrounded by blanks.

**Theorem 2.1.2.** *Given a monkey book, the wfd (word frequency distribution),* $P(n) = \frac{K(n)}{k}$*, where* $K(n)$ *is the the number of words which occur* $n$ *times and* $k$ *is the total number of words, in the continuum limit, is a power law. Denoting the wfd by* $p(n)$*, in the monkey book it is given by*

$$p(n) \propto n^{-\zeta} \tag{2.1}$$

*with*

$$\zeta = \frac{2\ln A - \ln(1-q_s)}{\ln A - \ln(1-q_s)}. \tag{2.2}$$

*Proof.* This monkey text has a certain information content given by the entropy of the letter configurations produced by the monkey. These configurations result in wfd $P(n)$ and the corresponding entropy $S = -\sum_n P(n)\ln P(n)$ gives a measure of the information associated with this frequency distribution. The most likely $P(n)$ corresponds to the maximum of $S$ under the appropriate constraints. This can equivalently be viewed as the minimum information loss, or cost, in comparison with an unconstrained $P(n)$. Consequently, the minimum cost $P(n)$ gives the most likely wfd for a monkey.

Let $n$ be the frequency with which a specific word occurs in a text and let the corresponding probability distribution be $p(n)dn$. This means that $p(n)dn$ is the probability that a word belongs to the frequency interval $[n, n+dn]$. The entropy associated with the probability distribution $p(n)$ is $S = -\sum_n p(n)\ln p(n)$ ($\sum$ implies an integral whenever the index is a continuous variable). Let $M(l)dl$ be the number of words in the word-letter length interval $[l, l+dl]$. This means that the number of words in the frequency interval $[n, n+dn]$ is $M(l)\frac{dl}{dn}dn$ in the degeneracy of a word with frequency $n$. The number of distinct words in the same interval is $K(n)dn = kp(n)dn$, which means that $\frac{M(l)}{K(n)}\frac{dl}{dn}$ is the degeneracy of a word with frequency $n$. The information loss due to this degeneracy is $\ln\left(\frac{M(l)}{K(n)}\frac{dl}{dn}\right) = \ln\left(M(l)\frac{dl}{dn}\right) - \ln p(n) + \text{const}$. So the average information

loss is given by

$$I_{cost} = \sum p(n) \left[ -\ln p(n) + \ln \left( M(l) \frac{dl}{dn} \right) \right] \tag{2.3}$$

and this is the appropriate information cost associated with the words: the $p(n)$ which minimized this cost corresponds to the most likely $p(n)$. The next step is to express $M(l)$ and $\frac{dl}{dn}$ in terms of the two basic probability distributions, $p(n)$ and the probability for hitting the keys. $M(l)$ is just $M(l) \approx A^l$. The frequency $n$ for a word containing $l$ letters is

$$n \approx \left( \frac{1 - q_s}{A} \right)^l q_s. \tag{2.4}$$

Thus $n \approx \exp(al)$ with $a = \ln(1 - q_s) - \ln A$ so that $\frac{dn}{dl} = na$ and, consequently,

$$I_{loss} = -\sum_n p(n) \ln p(n) + \sum_n p(n) \left[ \ln A^l - \ln na \right]. \tag{2.5}$$

Furthermore, $\ln \left( \frac{A^l}{na} \right) = l \ln A - \ln n - \ln a$ and from Eq. (2.4) we get $l = \frac{\ln \frac{n}{q_s}}{\ln \frac{1-q_s}{A}}$, from which follows that

$$\ln \frac{A^l}{na} = \left( -1 + \frac{\ln A}{\ln(1 - q_s) - \ln A} \right) \ln n + \text{const.}$$

Thus the most likely distribution corresponds to the minimum of the information word cost

$$I_{cost} = -\sum_n p(n) \ln p(n) + \sum_n p(n) \ln n^{-\zeta} \tag{2.6}$$

with

$$\zeta = \frac{2 \ln A - \ln(1 - q_s)}{\ln A - \ln(1 - q_s)}. \tag{2.7}$$

□

Morover, it can be easily proved that even Heaps' law is valid for monkey book.

**Theorem 2.1.3.** *Monkey books exhibit Heaps' law with the tyical relation* $\gamma = \frac{1}{z}$.

*Proof.* Suppose that a book of size $k$ has a wfd $P_k(n)$ created by sampling a fixed theoretical probability distribution $p(n) \propto n^{-\zeta}$, like a monkey book, where the normalization constant in only weakly dependent on $k$. The number of different words, $N_k$, for a given size is then related to $k$ through the relation

$$k = N_k \sum_{n=1}^{k} np(n) \tag{2.8}$$

and, since in the present case

$$\sum_{n=1}^{k} np(n) \propto \frac{1}{2-\zeta} \left( M^{2-\zeta} - 1 \right), \tag{2.9}$$

it follows that

$$N_k \propto k^{\zeta - 1} = k^{\frac{1}{z}}. \tag{2.10}$$

□

We have now proved that a monkey text exhibit both Zipf's and Heaps' laws and all relations between thier coefficients are valid. In order to confirm what we have up to now proved, I will show the results of some experiments I made with monkey books.

I wrote a program which, using different values of $A$ and fixed $q_s = 0.2$, creates monkey texts and corresponding Zipf's and Heaps' plots. First of all, we can observe, in the following figures, the behavior of Zipf's and Heaps' laws for two cases studied in my analysis.

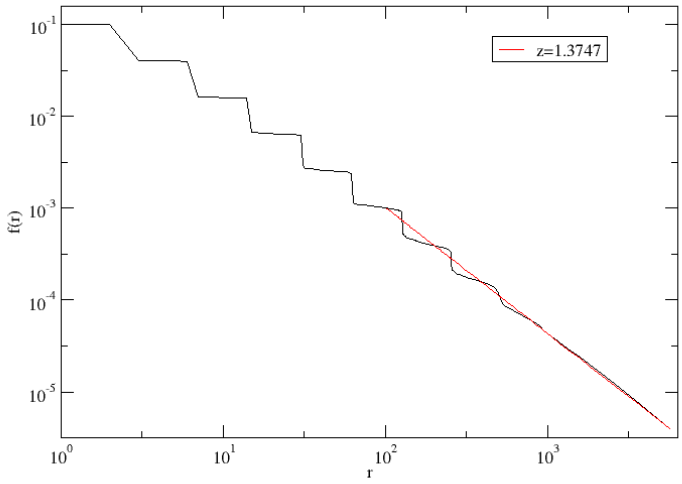



Figure 2.7: Zipf's law, in a $\log - \log$ plot, for a monkey text of 250000 words, with $A = 2$ and $q_s = 0.2$. $z = 1.3747 \pm 0.31323 \times 10^{-3}$

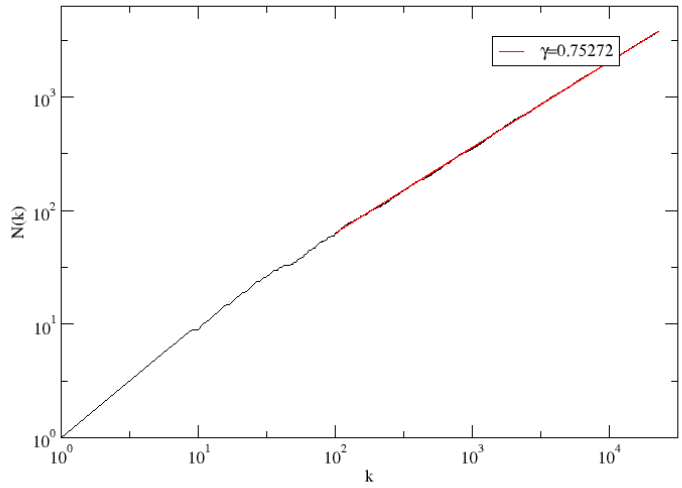



Figure 2.8: Heaps' law, in a $\log - \log$ plot, for a monkey text of 250000 words, with $A = 2$ and $q_s = 0.2$. $\gamma = 0.75272 \pm 0.50854 \times 10^{-6}$

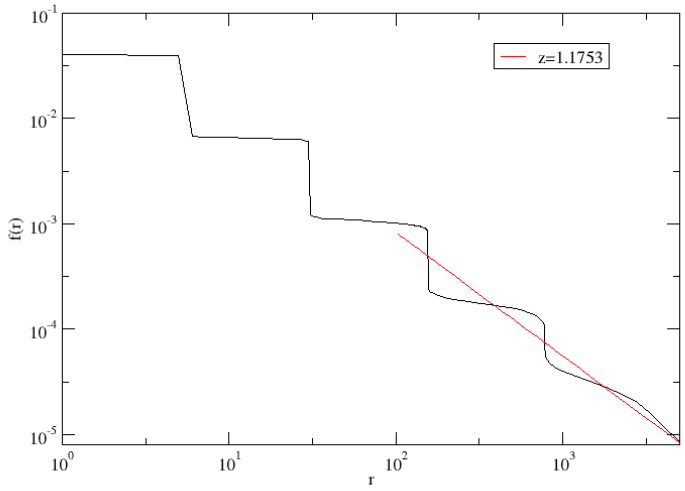



Figure 2.9: Zipf's law, in a $\log - \log$ plot, for a monkey text of 250000 words, with $A = 5$ and $q_s = 0.2$. $z = 1.1753 \pm 0.41098 \times 10^{-3}$

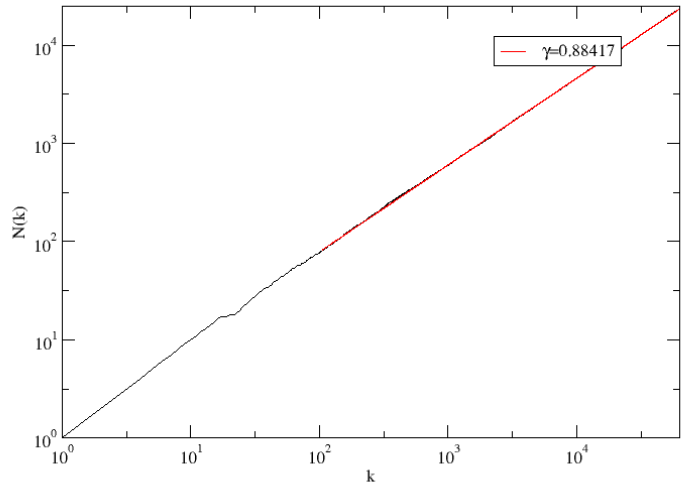



Figure 2.10: Heaps' law, in a $\log - \log$ plot, for a monkey text of 250000 words, with $A = 5$ and $q_s = 0.2$. $\gamma = 0.88417 \pm 0.13048 \times 10^{-6}$

As predicted, both Zipf's and Heaps' laws are valid in monkey books and we will see in the following tab that their coefficient follow all rules proved:

$$\begin{cases} \zeta_A^{ex} = \frac{2\ln A - \ln(1-q_s)}{\ln A - \ln(1-q_s)} \\ z_A^{ex} = \frac{1}{\zeta_A^{ex}-1} \\ \gamma_A^{ex} = \zeta_A^{ex} - 1 \end{cases} \tag{2.11}$$

| $A$ | $\zeta_A^{ex}$ | $z_A^{ex}$ | $\gamma_A^{ex}$ | $z$ | $\gamma$ |
|---|---|---|---|---|---|
| 2 | 1.756471 | 1.321928 | 0.756471 | 1.3721 | 0.78246 |
| 2 | 1.756471 | 1.321928 | 0.756471 | 1.3747 | 0.75272 |
| 2 | 1.756471 | 1.321928 | 0.756471 | 1.3849 | 0.76465 |
| 2 | 1.756471 | 1.321928 | 0.756471 | 1.3842 | 0.76075 |
| 2(average) | 1.756471 | 1.321928 | 0.756471 | 1.378975 | 0.765145 |
| 3 | 1.831176 | 1.203114 | 0.831176 | 1.2341 | 0.82561 |
| 3 | 1.831176 | 1.203114 | 0.831176 | 1.2462 | 0.82785 |
| 3 | 1.831176 | 1.203114 | 0.831176 | 1.2453 | 0.82359 |
| 3 | 1.831176 | 1.203114 | 0.831176 | 1.2456 | 0.83186 |
| 3(average) | 1.831176 | 1.203114 | 0.831176 | 1.2428 | 0.8272275 |
| 4 | 1.861353 | 1.160964 | 0.861353 | 1.1863 | 0.85702 |
| 4 | 1.861353 | 1.160964 | 0.861353 | 1.1833 | 0.85461 |
| 4 | 1.861353 | 1.160964 | 0.861353 | 1.1867 | 0.86414 |
| 4 | 1.861353 | 1.160964 | 0.861353 | 1.1858 | 0.86597 |
| 4(average) | 1.861353 | 1.160964 | 0.861353 | 1.185525 | 0.860435 |
| 5 | 1.878335 | 1.138647 | 0.878335 | 1.1753 | 0.88417 |
| 5 | 1.878335 | 1.138647 | 0.878335 | 1.1614 | 0.88338 |
| 5 | 1.878335 | 1.138647 | 0.878335 | 1.1689 | 0.87543 |
| 5 | 1.878335 | 1.138647 | 0.878335 | 1.1736 | 0.87786 |
| 5(average) | 1.878335 | 1.138647 | 0.878335 | 1.1698 | 0.88021 |

| $A$ | $\zeta_A^{ex}$ | $z_A^{ex}$ | $\gamma_A^{ex}$ | $z$ | $\gamma$ |
|---|---|---|---|---|---|
| 6 | 1.889253 | 1.124539 | 0.889253 | 1.1184 | 0.89472 |
| 6 | 1.889253 | 1.124539 | 0.889253 | 1.1201 | 0.89451 |
| 6 | 1.889253 | 1.124539 | 0.889253 | 1.1177 | 0.88443 |
| 6 | 1.889253 | 1.124539 | 0.889253 | 1.1211 | 0.89082 |
| 6(average) | 1.889253 | 1.124539 | 0.889253 | 1.119325 | 0.89112 |

Now, for concluding our study on monkey books, we can observe the following figures, where Zipf's and Heaps' plots are shown for all random texts created (4 texts for each $A$).

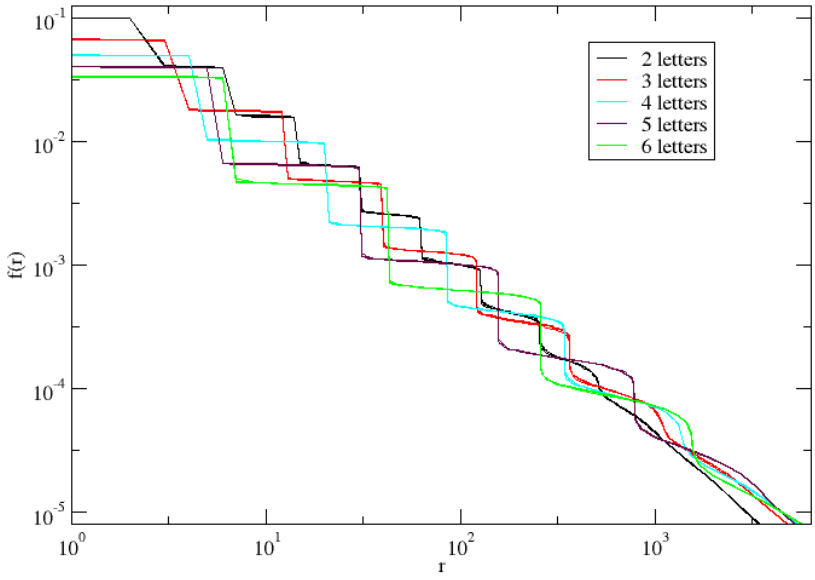



Figure 2.11: Zipf's law, in a $\log - \log$ plot, for all monkey texts of 250000 words, with $q_s = 0.2$

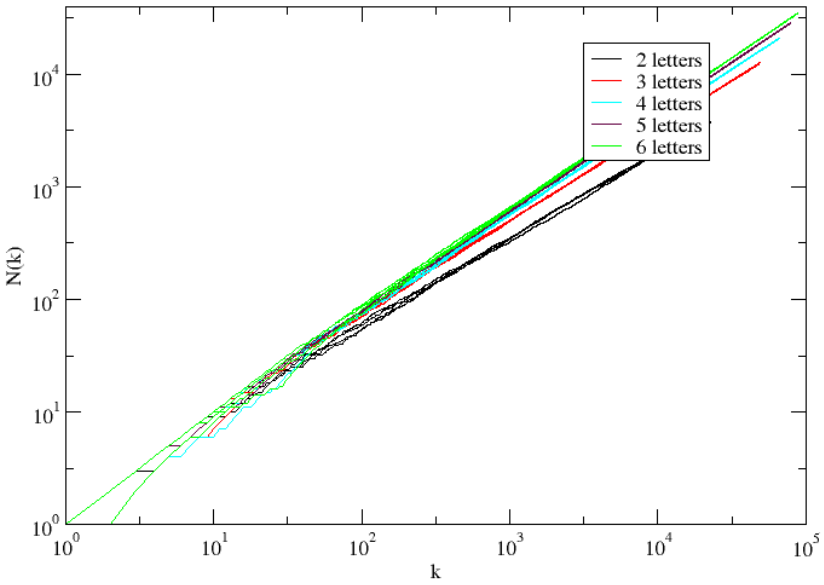



Figure 2.12: Heaps' law, in a $\log - \log$ plot, for all monkey texts of 250000 words, with $q_s = 0.2$

## 2.2 Real texts

In the previous chapter and section we have studied some models for the creation of texts which exhibit Zipf's and Heaps' laws, but, as I said before, these are just empirical observations and not laws in a rigorous sense. In this section we will observe Zipf's and Heaps' laws on real data, using the text *War and Peace* by Leo Tolstoj. Moreover we will analyze and study analogies and differences of these laws between different languages, using different translations of the same book in English, French, German and Italian.

Before beginning our Zipf's and Heaps' analysis, we should make a preliminary study on general informations about these different translations. In the following tab there are some basic statistical information about these texts.

| Language | English | French | German | Italian |
|---|---|---|---|---|
| Tot. n. characters | 3086648 | 2789763 | 3602335 | 3458573 |
| Tot. n. words | 572625 | 505476 | 582729 | 583357 |
| N. different words | 17543 | 21455 | 33202 | 31169 |
| $\frac{\text{N. different words}}{\text{Tot. n. words}}$ | 0.030636 | 0.042445 | 0.056977 | 0.053430 |
| Length sentences (avg.) | 19.474003 | 23.555870 | 21.562566 | 18.717179 |
| Length sentences (st.dev.) | 16.900486 | 21.573885 | 16.978340 | 17.892272 |
| Length words (avg.) | 4.390364 | 4.519087 | 5.181843 | 4.928764 |
| Length words (st.dev.) | 2.326076 | 2.768519 | 2.801607 | 2.870766 |

As predictable, although these values are of the same book in different translations, they are strongly dependent on language. Now we will see if even Zipf's and Heaps' laws depend on language.

First of all, in the following figures there are Zipf's and Heaps' plots for all studied languages.

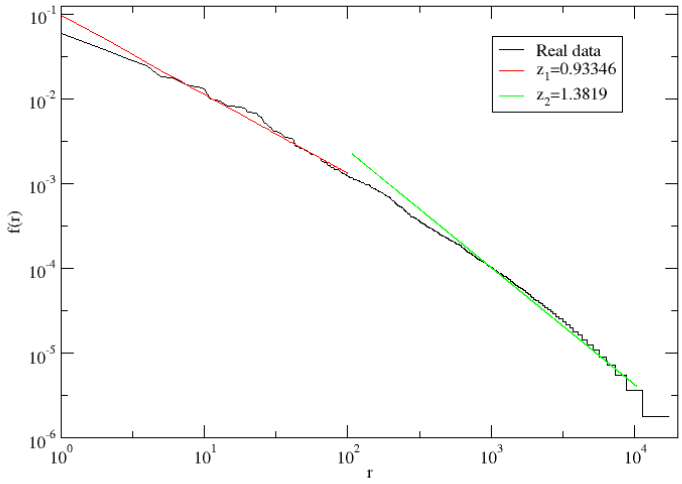



Figure 2.13: Zipf's law in a log-log plot for the book *War and Peace* in English. $z_1 = 0.93346 \pm 0.011811$, $z_2 = 1.3819 \pm 1.6770 \times 10^{-3}$.

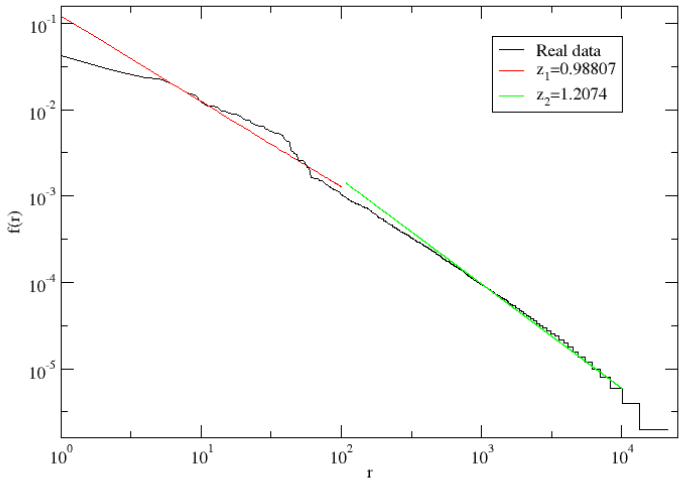



Figure 2.14: Zipf's law in a log-log plot for the book *War and Peace* in French. $z_1 = 0.98807 \pm 0.027223$, $z_2 = 1.2074 \pm 1.0761 \times 10^{-3}$.

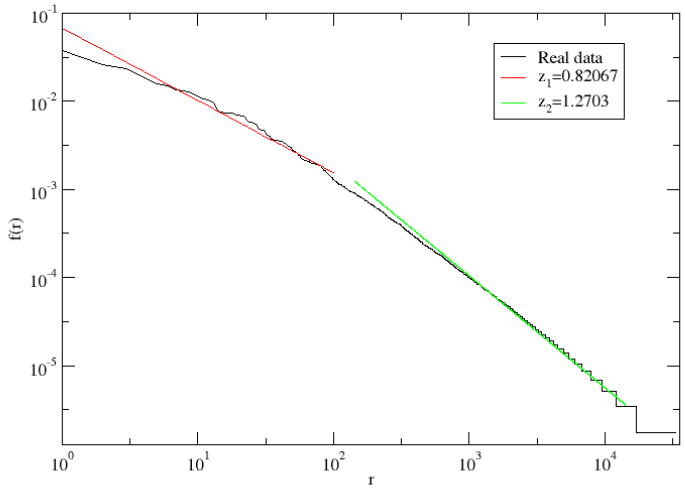



Figure 2.15: Zipf's law in a log-log plot for the book *War and Peace* in German. $z_1 = 0.82067 \pm 0.013216$, $z_2 = 1.2703 \pm 0.86616 \times 10^{-3}$.

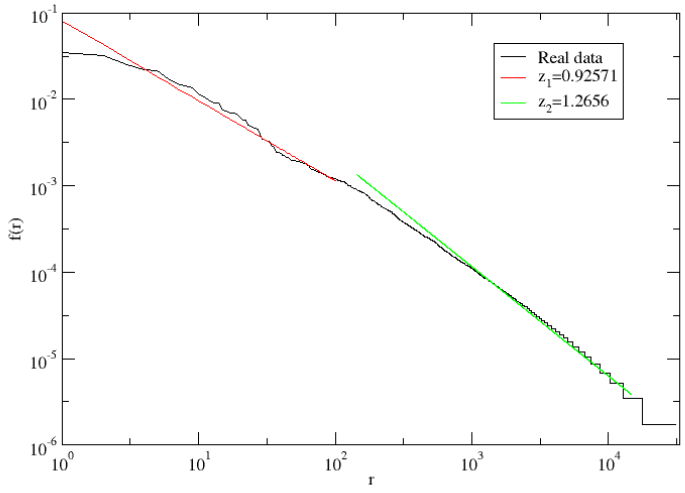



Figure 2.16: Zipf's law in a log-log plot for the book *War and Peace* in Italian. $z_1 = 0.92571 \pm 0.014773$, $z_2 = 1.2656 \pm 1.0137 \times 10^{-3}$.

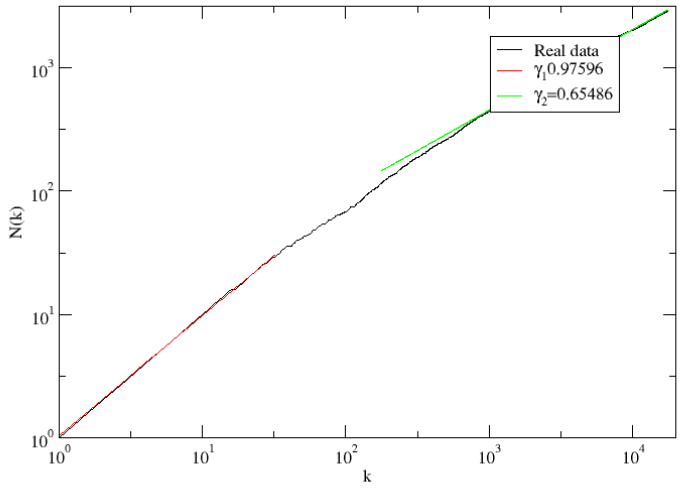



Figure 2.17: Heaps' law in a log-log plot for the book *War and Peace* in English. $\gamma_1 = 0.97596 \pm 4.5446 \times 10^{-3}$, $\gamma_2 = 0.65486 \pm 0.28741 \times 10^{-3}$.

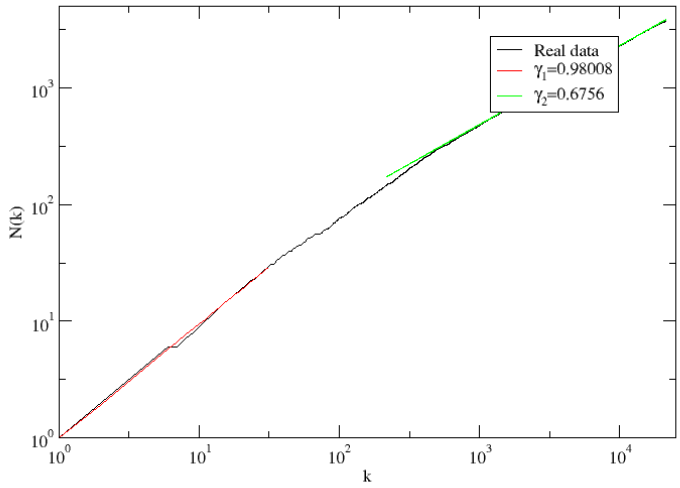



Figure 2.18: Heaps' law in a log-log plot for the book *War and Peace* in French. $\gamma_1 = 0.98008 \pm 8.0107 \times 10^{-3}$, $\gamma_2 = 0.6756 \pm 0.15711 \times 10^{-3}$.

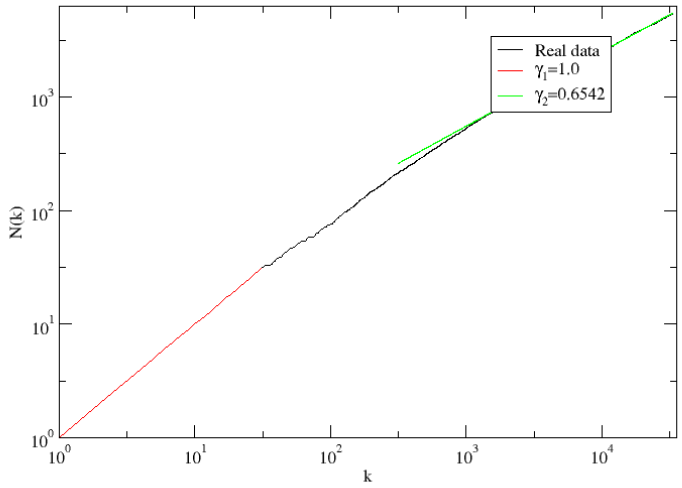



Figure 2.19: Heaps' law in a log-log plot for the book *War and Peace* in German. $\gamma_1 = 1.0 \pm 0$, $\gamma_2 = 0.6542 \pm 0.14241 \times 10^{-3}$.

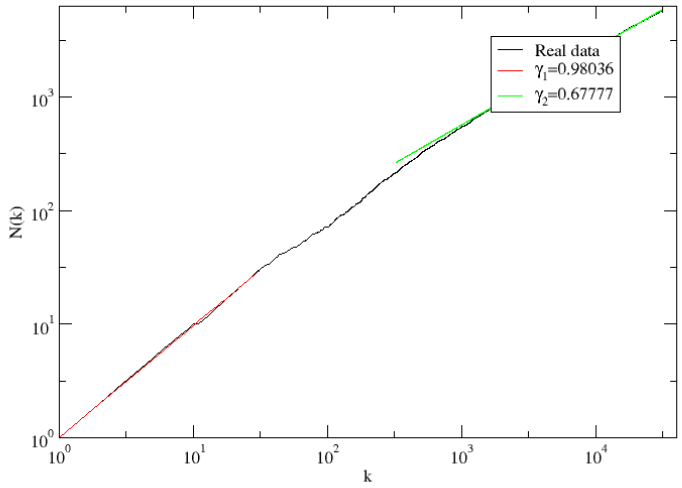



Figure 2.20: Heaps' law in a log-log plot for the book *War and Peace* in Italian. $\gamma_1 = 0.98036 \pm 5.4577 \times 10^{-3}$, $\gamma_2 = 0.67777 \pm 0.14564 \times 10^{-3}$.

In the following tab there is a summary of these results.

| Language | $z_1$ | $z_2$ | $\gamma_1$ | $\gamma_2$ | $\gamma_2^{ex}$ |
|---|---|---|---|---|---|
| English | 0.93346 | 1.3819 | 0.96742 | 0.655366 | 0.723641 |
| French | 0.98807 | 1.2074 | 0.98008 | 0.6756 | 0.828226 |
| German | 0.82067 | 1.2703 | 1.0 | 0.6542 | 0.787215 |
| Italian | 0.92571 | 1.2656 | 0.98036 | 0.67777 | 0.790139 |

where $\gamma_2^{ex} = \dfrac{1}{z_2}$ is the expected value of $\gamma_2$ as in Eq. (1.47).

Observing these figures and this tab, we can immediately note how both Zipf's and Heaps' laws follow a double power law, as previously exposed in Eqs. (1.5) and (1.18). In fact $z_1$ and $\gamma_1$ are both $\sim 1$ for all languages (except for the case of German language in Zipf's law) and there is a "turning point" in which the exponents change. We even have to note how $\gamma_2 \not\approx \gamma_2^{ex}$ but this is probably due to the fact that in this experiment we don't use a book long enough to describe an asymptotic behavior.

Another fundamental observation we can do is that Zipf's and especially Heaps' asymptotic coefficients don't show a strong dependence on the chosen language and the only couple of values which deviate consistently from the other values are $z_2$ for English and French. This result was anyway unexpected, in fact different languages are so different from lexical and grammatical point of view (for example in German and Italian there are many words which are combination of other words) that we expected that, even if they follow the same laws, their coefficients were different. Anyway it should be very interesting to continue the investigation on this particular result, maybe using a bigger corpus and other languages belonging to other families of language (all languages I used for my study are Indio-European languages), like Japanese, Russian, Chinese and so on.

In these first two chapter we have shown that, not only real texts, but also many texts created using various models, exhibit Zipf's and Heaps' laws, and it should be very interesting to study differences and similarities between Zipf's and Heaps' laws in texts created using various models and real texts. This particular research was introduced by R. F. i Cancho and B. Elevåg [40]. They compared Zipf's law in real texts and in monkey books, obtaining the result that, using different values of $A$ and $q_s$, Zipf's law obtained is extremely different from Zipf's law exhibited in real texts. This result can be easily observed in the following figure, caught from [40], in which there are Zipf's plots for real texts and for monkey books with parameters caught from real texts ($q_s$ is the average of word length in real texts and $A = 26$).

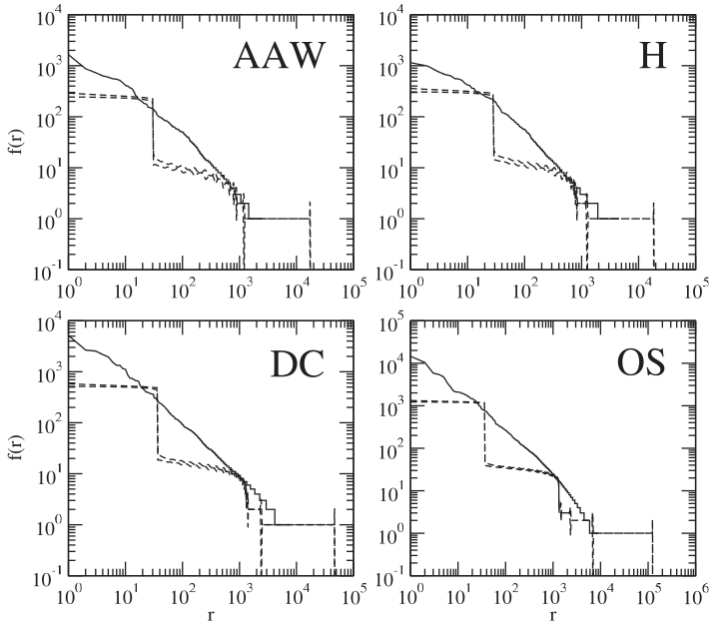



Figure 2.21: Differences between Zipf's law in real texts (thin black line) and two control curves of the expected histogram of a monkey text of the same length in words (dashed lines) involving four English texts, *Alice's Adventures in Wonderland* (AAW), *Hamlet* (H), *David Crockett* (DC) and *The Origin of Species* (OS). $f(r)$ is the number of occurrences of a word of rank $r$.

R. F. i Cancho and B. Elevåg began this research and it should be interesting to continue following this basic idea.

Another interesting question to be answered is the following: how Zipf's and Heaps' exponents of various models depend on features of language? First of all we should check if parameters of different models depend on language and how these parameters influence Zipf's and Heaps' exponents. In the following tab there is exactly this preliminary analysis for models analyzed above. LEAST is the model proposed by R. F. i Cancho and R. V. Solè [27] and explained and analyzed in **1.1.2**; SIMON is Simon's model [3] explained and analyzed in **1.3.1** and in **2.1.1**; GE.ALT. is Gerlach's and Altmann's model [45] explained and analyzed in **1.3.2**; MONKEY is the model for creating a monkey text [41] explained and analyzed in **2.1.2**. Note that in the following tab, in the third column there is a ✓if and only if there is a ✓in both other columns.

| Model | Parameters depend on language? | Exponents depend on parameters? | Exponents depend on language? |
|---|---|---|---|
| LEAST | × | × | × |
| SIMON | ✓ | ✓ | ✓ |
| GE.ALT. | ✓ | ✓ | ✓ |
| MONKEY | × | ✓ | × |

| Model | $z$ | $\gamma$ |
|---|---|---|
| LEAST | 1 | ? |
| SIMON | $1-\alpha$ | 1 |
| GE.ALT. | $1+\alpha$ | $(1+\alpha)^{-1}$ |
| MONKEY | $1-\frac{\ln(1-q_s)}{\ln A}$ | $\left(1-\frac{\ln(1-q_s)}{\ln A}\right)^{-1}$ |

Moreover, from subsection **1.3.1** and section **2.1** we can easily note how Simon's model's results are inconsistent with results obtained for real texts. In fact $\gamma_{\text{simon}} = 1$ and $z_{\text{simon}} < 1$, while for real texts $z, \gamma > 1$.

Hence it may be interesting to do some experiments on Gerlach's and Altmann's model, the most complete one, in order to check if results, obtained with different values of parameters dependent on different languages, are able to consistently explain the differences in Zipf's and Heaps' exponents in real texts.

# Chapter 3

# Long-Range Correlations and Burstiness

In the previous chapters we have studied two linguistic laws, that detect some global statistical features of texts. Now we will concentrate our study on two other features that, on the contrary, study particular properties of chosen words or conditions: burstiness and long-range correlations. For this chapter I mainly used [43] by E. G. Altmann, G. Cristadoro and M. Degli Esposti.

## 3.1 Introduction

Following the information theory approach, we consider a literary text as the output of a stationary and ergodic source that takes values in a finite alphabet and we search information about the source through a statistical analysis of the text.

In this section we will mainly focus our study on correlation functions, which are defined after specifying an observable and a product over the defined observables. Given a symbolic sequence $s$, we denote by $s_k$ the symbol in the $k$th position and by $s_n^m$ $(m \geq n)$ the subsequence $(s_n, s_{n+1}, ..., s_m)$.

We define observables as functions $f$ that map symbolic sequences $s$ into a number sequence $x$ (we will use binary sequences of 0 and 1). If we focus on local mappings, we define $x_k = f\left(s_k^{k+r}\right)$, for a fixed $r \geq 0$ and any $k$. Its autocorrelation function is defined as:

$$C_f(t) = \langle f\left(s_i^{i+r}\right) f\left(s_{i+t}^{i+t+r}\right)\rangle - \langle f\left(s_i^{i+r}\right)\rangle\langle f\left(s_{i+t}^{i+t+r}\right)\rangle \tag{3.1}$$

where $t$ plays the role of the time (counted in numbers of symbols) and $\langle . \rangle$ denotes an average over sliding windows. In order to better understand this definition we have to analyze the previous equation.

*Remark* 3.1.1. Given an ergodic and stationary process, correlation functions are defined as

$$\mathrm{Corr}_x(j,t) = \mathbb{E}(x_j x_{j+t}) - \mathbb{E}(x_j)\mathbb{E}(x_{j+t}) \tag{3.2}$$

where $\mathbb{E}(.)$ is an average over different realizations $x$ of the process. Stationarity guarantees that $\mathrm{Corr}(j,t)$ depends on time lag $t$ only. In our case, any binary sequence $x$ is obtained from a single text of length $N$ using a given map. In such cases it is possible to assume ergodicity to approximate the Eq. (3.2) by

$$C_x(t) = \langle x_j x_{j+t}\rangle - \langle x_j\rangle\langle x_{j+t}\rangle \tag{3.3}$$

where $\langle . \rangle$ means the average, for each fixed $t$, over all pairs $x_j$ and $x_{j+t}$ as

$$\langle . \rangle \equiv \frac{1}{N-t}\sum_{j=1}^{N-t} . \tag{3.4}$$

Now, the choice of the observable $f$ is fundamental to determine which "memory" of the source we want to quantify. Only once a class of observables with the same properties shows the asymptotic autocorrelation, it is possible to think about long-range correlations of the text as a whole. The observable we will use is the following.

**Definition 3.1.1.**

$$x_k = f_\alpha(s_k) = \begin{cases} 1, & \text{if condition } \alpha \text{ is verified,} \\ 0, & \text{if condition } \alpha \text{ is not verified.} \end{cases} \tag{3.5}$$

Once obtained the binary sequence $x$ associated to the chosen condition $\alpha$ we can study the asymptotic behavior of its $C_x(t)$. We are particularly interested in the long-range correlated case

$$C_x(t) = \langle x_j x_{j+t} \rangle - \langle x_j \rangle \langle x_{j+t} \rangle \propto t^{-\beta}, \quad 0 < \beta < 1, \tag{3.6}$$

for which $\sum_{t=0}^{\infty} C_x(t)$ diverges.

Before continuing our analysis we have to show two fundamental results, especially the second one, the **Theorem 3.1.2**, for the study and the implementation of algorithms we will use next.
The following theorem is caught from [18].

**Theorem 3.1.1.** *In the long-range correlated case explained above, the power spectrum, defined as* $S(f) = C_x(0) + 2\sum_{t=1}^{\infty} C_x(t)\cos(2\pi f t)$*, follows the following*

$$S(f) \propto f^{-\alpha} \tag{3.7}$$

*for small* $f$ *where* $\alpha = 1 - \beta$.

*Proof.* If $C_x(t)$ obeys the scaling relation in Eq. (3.6) then

$$S(f) \approx 2\sum_{t=1}^{\infty} t^{-\beta}\cos(2\pi f t). \tag{3.8}$$

Consider the Taylor expansion of the function $(1-y)^{-\delta-1}$,

$$(1-y)^{-\delta-1} = \sum_{t=0}^{\infty} A_t^{\delta} y^t, \tag{3.9}$$

where by definition we have $A_0^{\delta} = 1$ and, for $t \geq 1$,

$$A_t^{\delta} = \frac{(\delta+1)(\delta+2)...(\delta+t)}{t!} \approx \frac{t^{\delta}}{\Gamma(\delta+1)}. \tag{3.10}$$

This means

$$\sum_{t=1}^{\infty} t^{\delta} y^t \approx \Gamma(\delta+1)\left[(1-y)^{-\delta-1} - 1\right]. \tag{3.11}$$

Replacing $\delta = -\beta$, $y = r\exp(i2\pi f)$ and $0 \leq r < 1$ in the above equation leads to

$$\sum_{t=1}^{\infty} t^{-\beta} r^t \exp(i2\pi tf) \approx \Gamma(1-\beta)\left[(1 - r\exp(i2\pi f)^{\beta-1} - 1\right]. \tag{3.12}$$

Letting $r \to 1$ and $f \to 0$, and taking the real part, we obtain

$$\sum_{t=1}^{\infty} t^{-\beta} \cos(2\pi tf) \approx \Gamma(1-\beta)(2\pi f)^{\beta-1} \cos\left[\frac{\pi}{2}(1-\beta)\right] \tag{3.13}$$

Substituting this into Eq. (3.8) yields

$$S(f) \approx 2\Gamma(1-\beta)(2\pi f)^{\beta-1} \cos\left[\frac{\pi}{2}(1-\beta)\right] \propto f^{\beta-1} = f^{-\alpha} \tag{3.14}$$

□

**Theorem 3.1.2.** *In the long-range correlated case explained above, the associate random walker* $X(t) = \sum_{j=0}^{t} x_j$ *spreads super-diffusively as*

$$\sigma_X^2(t) = \langle X(t)^2\rangle - \langle X(t)\rangle^2 \propto t^{\gamma} \tag{3.15}$$

*where* $\gamma = 2 - \beta$.

*Proof.*

$$\langle X(t)^2\rangle = \sum_{i=1}^{t}\langle x_i^2\rangle + 2\sum_{s=1}^{t-1}(t-s)\langle x_i x_{i+s}\rangle =$$

$$= t\langle x_i^2\rangle + 2t\sum_{s=1}^{t-1} C_x(s) - 2\sum_{s=1}^{t-1} sC_x(s) + 2t\sum_{s=1}^{t-1}\langle x_i\rangle\langle x_{i+s}\rangle - 2\sum_{s=1}^{t-1} s\langle x_i\rangle\langle x_{i+s}\rangle =$$

$$= t\langle x_i^2\rangle + 2t\sum_{s=1}^{t-1} C_x(s) - 2\sum_{s=1}^{t-1} sC_x(s) + 2t(t-1)\langle x_i\rangle^2 - 2\frac{t(t-1)}{2}\langle x_i\rangle^2 =$$

$$= t\langle x_i^2\rangle + 2t\sum_{s=1}^{t-1} C_x(s) - 2\sum_{s=1}^{t-1} sC_x(s) + t(t-1)\langle x_i\rangle^2 \tag{3.16}$$

while

$$\langle X(t)\rangle = t\langle x_i\rangle. \tag{3.17}$$

Thus

$$\sigma_X^2(t) = \langle X(t)^2\rangle - \langle X(t)\rangle^2 =$$
$$= t\langle x_i^2\rangle + 2t\sum_{s=1}^{t-1} C_x(s) - 2\sum_{s=1}^{t-1} sC_x(s) + t(t-1)\langle x_i\rangle^2 - t^2\langle x_i\rangle^2 =$$
$$= t\langle x_i^2\rangle + 2t\sum_{s=1}^{t-1} C_x(s) - 2\sum_{s=1}^{t-1} sC_x(s) - t\langle x_i\rangle^2 =$$
$$= 2t\sum_{s=1}^{t-1} C_x(s) - 2\sum_{s=1}^{t-1} sC_x(s) + t\langle x_i\rangle\left(1 - \langle x_i\rangle\right) \tag{3.18}$$

Now let $C_x(s)$ obey the scaling law in Eq.(3.6). The sums in the above equation are estimated as

$$\sum_{s=1}^{t-1} C_x(s) \propto \sum_{s=1}^{t} s^{-\beta} \approx \int_1^t s^{-\beta} = t^{1-\beta} \tag{3.19}$$

and

$$\sum_{s=1}^{t-1} sC_x(s) \propto \sum_{s=1}^{t} s^{1-\beta} \approx \int_1^t s^{1-\beta} = t^{2-\beta} \tag{3.20}$$

For $0 < \beta < 1$, this means

$$\sigma_X^2(t) = \langle X(t)^2\rangle - \langle X(t)\rangle^2 \propto t^{2-\beta} \tag{3.21}$$

for large $t$.

□

Our study on the possible origins of the long-range correlations will be based on the relation between the power spectrum $S(f)$ at $f = 0$ and the statistics of the sequence of inter-event times $\tau_i$'s (one plus the numbers of zeros between two consecutive ones). For the short range correlated case, $S(0)$ is finite and given by:

$$S(0) = \frac{\sigma_\tau^2}{\langle\tau\rangle^3}\left(1 + 2\sum_k C_\tau(k)\right). \tag{3.22}$$

On the contrary, for the long-range correlated case $S(0) \to \infty$ and Eq. (3.22) implies two possible origins:

- *burstiness* measured as the broad tail of the distribution of inter-event times $p(\tau)$ (divergent $\sigma_\tau$);

- *long-range correlations* of the sequence of $\tau_i$'s (not summable $C_\tau(k)$).

## 3.2 Hierarchy of Natural Language

In this section we will introduce a hierarchy of language and study how moving from different hierarchical levels affects long-range correlations and burstiness.

### 3.2.1 Explanation of hierarchy

Our hierarchy is built in the following way. Levels are established from sets of semantically or syntactically conditions $\alpha$'s (for example vowels- consonants, different letters, different words, different topics). Each binary sequence $x$ is obtained by mapping the text using the observable $f_\alpha$ given above and will be indicated by the fixed condition $\alpha$. For example, *prince* indicates the sequence $x$ obtained from the condition $\alpha$ : $s_k^{k+8} =$ " *prince* ". A sequence $z$ is linked to $x$ if $\forall j$ such that $x_j = 1$ we have $z_{j+r'} = 1$ for a fixed constant $r'$. In such a case we say that $x$ is on top of $z$ and that $x$ belongs to a higher level than $z$. Obviously, there are no links between sequences at the same level and a sequence at a given levels is on top of all those sequences at lower levels, linked with a direct path. For example, *" prince "* is on top of *"e"*, *"e"* is on top of *vowels*, so *" prince "* is on top of *vowels*. As will be clear later from the results, the concept of link can be extended to a probabilistic meaning (for example *" prince "* is more probable to appear in a part whose topic is connected to war).

This hierarchy can be better understood thanks to the following figure, caught from [43].

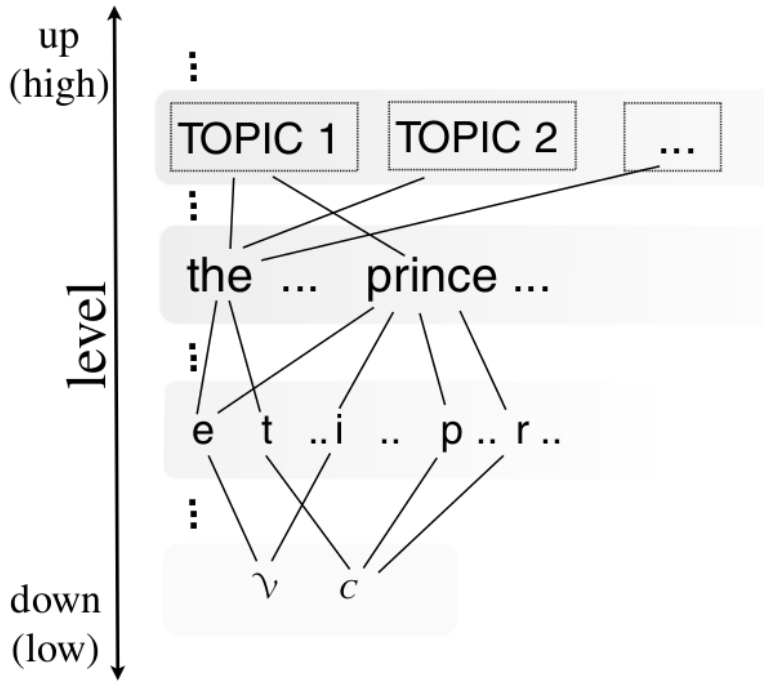



Figure 3.1: Hierarchy of levels at which literary texts can be analyzed. Depicted are the levels vowels-consonants ($\nu/c$), letters (a-z), words and topics.

### 3.2.2 Moving in the hierarchy

In this subsection we will show how correlations behave between two linked sequences. Let $x$ be a sequence on top of $z$ and $y$ be the unique sequence on top of $z$ such that $z = x + y$, that means $z_i = x_i + y_i,\ \forall i$. The spreading of the walker $Z$, associated to $z$, is

$$\sigma_Z^2(t) = \sigma_X^2(t) + \sigma_Y^2(t) + 2C\left(X(t), Y(t)\right), \tag{3.23}$$

where $C(A, B) = \langle AB\rangle - \langle A\rangle\langle B\rangle$.

**Lemma 3.2.1.** *If $z = x + y$, then*

$$\sigma_Z(t) \leq \sigma_X(t) + \sigma_Y(t). \tag{3.24}$$

*Proof.* Using the Cauchy-Schwarz inequality we know that $|C\left(X(t), Y(t)\right)| \leq \sigma_X(t)\sigma_Y(t)$. Using this in Eq. (3.23) we obtain $\sigma_Z(t) \leq \sigma_X(t) + \sigma_Y(t)$. □

Now defining $\overline{x}$ as the sequence obtained inverting $0 \leftrightarrow 1$ on each of its elements $\overline{x}_i = 1 - x_i$.

**Lemma 3.2.2.** *If $z = x + y$, then $\overline{x} = \overline{z} + y$ and $\overline{y} = \overline{z} + x$.*

*Proof.* We know that $z_i = x_i + y_i$, then $1 - \overline{z}_i = 1 - \overline{x}_i + y_i \ \Rightarrow \overline{x}_i = \overline{z}_i + y_i$. In an analogous way, $1 - \overline{z}_i = 1 - \overline{y}_i + x_i \ \Rightarrow \overline{y}_i = \overline{z}_i + x_i$ □

**Lemma 3.2.3.** $\sigma_{\overline{X}}(t) = \sigma_X(t)$.

*Proof.* $\overline{X}(t) = \sum_{i=1}^{t} 1 - x_i = t - \sum_{i=1}^{t} x_i = t - X(t)$. Then
$\sigma^2_{\overline{X}}(t) = \sigma^2(t) + \sigma^2(X(t)) = \sigma^2_X(t)$. □

**Theorem 3.2.4.** *If $z = x + y$, the following equations are valid*

$$\sigma_X(t) \leq \sigma_Z(t) + \sigma_Y(t) \tag{3.25}$$

$$\sigma_Y(t) \leq \sigma_Z(t) + \sigma_X(t). \tag{3.26}$$

*Proof.* We know from **Lemma 3.2.2** that

$$\overline{x} = \overline{z} + y \text{ and } \overline{y} = \overline{z} + x,$$

so, from **Lemma 3.2.1**, we obtain that

$$\sigma_{\overline{X}}(t) \leq \sigma_{\overline{Z}} + \sigma_Y(t) \text{ and } \sigma_{\overline{Y}}(t) \leq \sigma_{\overline{Z}} + \sigma_X(t).$$

Using **Lemma 3.2.3** we obtain

$$\sigma_X(t) \leq \sigma_Z(t) + \sigma_Y(t) \text{ and } \sigma_Y(t) \leq \sigma_Z(t) + \sigma_X(t).$$

□

If now we suppose that $\sigma_i^2(t) \propto t^{\gamma_i}$ with $i \in \{X, Y, Z\}$, then, in order to satisfy the above inequalities, at least two of the three $\gamma_i$ have to be equal to $\max_i \{\gamma_i\}$.

Now we discuss the implications of this result to the behavior of correlations moving up and down in the hierarchy.

**Up:** Supposing that at a given level we have a binary sequence $z$ with long-range correlation $\gamma_Z > 1$, then we know from the result explained above that there is at least a sequence $x$ on top of $z$ with long-range correlation $\gamma_X \geq \gamma_Z$. This implies, for example, that if we observe long-range correlations in the sequence associated to a letter then we can argue that its anomaly comes from the anomaly of at least a word, where this letter appears.

**Down:** Supposing that $x$ is long-range correlated $\gamma_X > 1$, then from Eq. (3.23) and from the result obtained above, in order to get $\gamma_Z < \gamma_X$, we should have $\gamma_X = \gamma_Y$, which is extremely unlikely in the typical case of sequences $z$, which receive contributions from different sources (for example a letter receives contribution from different words). So we can consider $z$ to be composed by $n$ sequences $x^{(j)}$, $j = 1, ..., n$ with $\gamma_{X^{(1)}} \neq \gamma_{X^{(2)}} \neq ... \neq \gamma_{X^{(n)}}$, obtaining $\gamma_Z = \max_j \{\gamma_{X^{(j)}}\}$. In conclusion, correlations typically flow down in our hierarchy of levels.

### 3.2.3 Finite time effects

While the results up to now shown are valid asymptotically (infinitely long sequences), in the case of a real text we only have a finite time estimate $\hat{\gamma}$ of the real value of $\gamma$. Always from Eq. (3.23) we can note that, adding sequences $x_j$ with different $\gamma_{X^{(j)}}$ and following the procedure used for moving down in the hierarchy, leads to $\hat{\gamma}_Z < \gamma_Z$ if $\hat{\gamma}_Z$ is computed at a time when the asymptotic regime is still not dominating. This fact, as we will see later, is a fundamental feature in the analysis of long-range correlations in real books. Now, in order to give quantitative estimates, we consider a sequence $z$ that is the sum of the most long-range correlated sequence $x$ (the one with $\gamma_X = \max_j \{\gamma_{X^{(j)}}\}$) and another independent non overlapping sequence ($y$ is non overlapping with $x$ if $x_i = 1 \Rightarrow y_i = 0$). So, defining $y = \xi(1 - x)$ with $\xi_i$ I.I.D. binary random variable, then $z = x + y$ corresponds to a random addition of ones with probability $\langle \xi \rangle$ to the zeros of $x$.

**Theorem 3.2.5.** *The associated random walker $Z$ spreads super-diffusively with the same exponent of $X$.*

*Proof.* As written in Eq. (3.23), we know that

$$\sigma_Z^2(t) = \sigma_X^2(t) + \sigma_Y^2(t) + 2C\left(X(t), Y(t)\right).$$

In this case we have

$$\langle Y(t)\rangle = \langle\xi\rangle t\left(1 - \langle x\rangle\right) \tag{3.27}$$

and

$$\langle Y(t)^2\rangle = \langle\sum_{i,j=1}^{t}(\overline{x}_i\xi_i)(\overline{x}_j\xi_j)\rangle = \langle\sum_{i=1}^{t}(\overline{x}_i^2\xi_i^2)\rangle + \langle\sum_{i,j=1;i\neq j}^{t}(\overline{x}_i\xi_i)(\overline{x}_j\xi_j)\rangle =$$

$$= \sum_{i=1}^{t}\langle\overline{x}_i^2\rangle\langle\xi^2\rangle + \sum_{i,j=1;i\neq j}^{t}\langle\overline{x}_i\overline{x}_j\rangle\langle\xi\rangle^2 =$$

$$= \sum_{i=1}^{t}\langle\overline{x}_i^2\rangle\langle\xi^2\rangle - \sum_{i=1}^{t}\langle\overline{x}_i^2\rangle\langle\xi\rangle^2 + \sum_{i=1}^{t}\langle\overline{x}_i^2\rangle\langle\xi\rangle^2 + \sum_{i,j=1;i\neq j}^{t}\langle\overline{x}_i\overline{x}_j\rangle\langle\xi\rangle^2 =$$

$$= \langle\xi\rangle^2\langle X(t)^2\rangle + \sigma^2(\xi)\sum_{i=1}^{t}\langle\overline{x}_i^2\rangle. \tag{3.28}$$

From Eqs. (3.27) and (3.28), from $\sum_{i=1}^{t}\langle\overline{x}_i^2\rangle = \sum_{i=1}^{t}\langle\overline{x}_i\rangle$ and from $\sigma_{\overline{X}}^2(t) = \sigma_X^2(t)$ we obtain

$$\sigma_Y^2(t) \equiv \langle Y(t)^2\rangle - \langle Y(t)\rangle^2 = \langle\xi\rangle^2\sigma_X^2(t) + t\sigma_\xi^2(1 - \langle x\rangle). \tag{3.29}$$

The correlation term in Eq.(3.23) can be directly calculated:

$$C(X(t), Y(t)) = \langle X(t)Y(t)\rangle - \langle X(t)\rangle\langle Y(t)\rangle =$$

$$= \langle\sum_{i,j=1}^{t}x_i(1 - x_j)\xi_j\rangle - \langle X(t)\rangle\langle\sum_{j=1}^{t}(1 - x_j)\xi_j\rangle =$$

$$= \langle X(t)\rangle\langle\xi\rangle t - \langle X(t)^2\rangle\langle\xi\rangle - \langle X(t)\rangle\left[\langle\xi\rangle t - \langle X(t)\rangle\langle\xi\rangle\right] =$$

$$= -\langle\xi\rangle\sigma_X^2(t). \tag{3.30}$$

Now, combining Eqs. (3.29), (3.30) and (3.23) we have

$$\sigma_Z^2(t) = \sigma_X^2(t) + \sigma_Y^2(t) + 2C\left(X(t), Y(t)\right) =$$

$$= \sigma_X^2(t) + \langle\xi\rangle^2\sigma_X^2(t) + t\sigma_\xi^2(1-\langle x\rangle) - 2\langle\xi\rangle\sigma_X^2(t) =$$

$$= t\sigma_\xi^2(1-\langle x\rangle) + \sigma_X^2(t)\left(1-\langle\xi\rangle\right)^2 =$$

$$= \langle\xi\rangle\left(1-\langle\xi\rangle\right)\left(1-\langle x\rangle\right)t + \left(1-\langle\xi\rangle\right)^2\sigma_X^2(t). \tag{3.31}$$

In conclusion, if $X$ spreads super-diffusively, then even $Z$ spreads super-diffusively too and they both have the same exponent and hence the asymptotic behavior. □

We even have to consider that the asymptotic regime is hidden at short times by a pre-asymptotic normal behavior, given by the linear term in $t$. We can even emphasize that, even if the condition for $y$ not to be overlapping forces both $\sigma_Y^2(t)$ and $C\left(X(t), Y(t)\right)$ to have the same asymptotic behavior of $\sigma_X^2(t)$, their cumulative contributions don't vanish unless $\langle\xi\rangle = 1$.

We can now give a bound on the transition time $t_T$ to the asymptotic diffusion exposed in Eq. (3.31). Consider the case in which even the asymptotic anomalous behavior of $X$ is hidden by a generic pre-asymptotic $A(t)$ such that

$$\sigma_X^2(t) = \langle x\rangle\left(1-\langle x\rangle\right)\left[(1-g)A(t) + gt^{\gamma_X}\right], \tag{3.32}$$

with $0 < g \leq 1$ and $A(t)$ increasing and such that $\dfrac{A(t)}{t^{\gamma_X}} \to 0$ for $t \to \infty$ and $A(1) = 1$. The first condition guarantees that the asymptotic behavior is dominated by $t^{\gamma_X}$, while the second one guarantees that $\sigma_X^2(1) = \langle x\rangle\left(1-\langle x\rangle\right)$.

The asymptotic behavior $\sigma_X^2(t) \propto t^{\gamma_X}$ in Eq. (3.31) dominates only after a time $t_T$ such that:

$$\frac{\langle\xi\rangle t_T + (1-g)\langle x\rangle\left(1-\langle\xi\rangle\right)A(t_T)}{g\left(1-\langle\xi\rangle\right)\langle x\rangle} = t_T^{\gamma_X}. \tag{3.33}$$

We even know that $(1-g)\langle x\rangle\,(1-\langle\xi\rangle)\,A(t_T) > 0$ and that $t^{\gamma_X}$ is monotonically increasing, so we finally have

$$t_T \geq t_T^* = \left(\frac{\langle\xi\rangle}{(1-\langle\xi\rangle)\,g\langle x\rangle}\right)^{\frac{1}{\gamma_X{}^{-1}}} \tag{3.34}$$

In conclusion, any finite time estimate $\hat{\gamma}_X$ is close to the real asymptotic $\gamma_X$ only if the estimate is performed for $t \gg t_T$, otherwise $\hat{\gamma}_X < \gamma_X$ and $\hat{\gamma}_X \approx 1$ if $t \ll t_T$.

As noted before in **Lemma 3.2.2**, if $z = x + y$ then $\overline{x} = \overline{z} + y$. Applying the procedure used above to this relation, similar pre-asymptotic normal diffusion and transition time appear in the case of random subtraction, moving up in the hierarchy. In practice, starting from a sequence $z$ that asymptotically behaves as $\sigma_Z^2(t) \backsimeq g\langle z\rangle\,(1-\langle z\rangle)\,t^{\gamma_Z}$ and constructing $x = \zeta z$, with $\zeta$ a binary sequence independent from $z$, we obtain a transition time $t_T$ for $x$ given by:

$$t_T \geq t_T^* = \left(\frac{1-\langle\xi\rangle}{(1-\langle z\rangle)\,g\langle\xi\rangle}\right)^{\frac{1}{\gamma_Z{}^{-1}}}, \tag{3.35}$$

which corresponds to Eq. (3.34) after properly replacing $\langle x\rangle \to (1-\langle z\rangle)$, $\langle\xi\rangle \to (1-\langle\xi\rangle)$ and $\gamma_X \to \gamma_Z$.

In contrast with correlations, burstiness, due to the tails of the inter-event time distribution $p(\tau)$, isn't preserved through a movement up and down in the hierarchy. If we consider going down by adding sequences with different tails of $p(\tau)$, then the tail of this new sequence will be bound by the shortest tail of the individual sequences. Considering the random addition example, explained above, $z = x + \xi(1-x)$ where $x$ has a broad tail in $p(\tau)$, then $p(\tau)$ for $z$ has short tails because the cluster of zeros in $x$ is cut randomly by $\xi$. Going up in the hierarchy, if we take a sequence on top of a given bursty binary sequence, for example using the random subtraction explained above $x = \zeta z$, then the probability of finding a large inter-event time $\tau$ in $z$ increases as the number of times the random elimination joins two or more

clusters of zeros in $x$, and decreases as the number of times the elimination destroys a pre-existent inter-event time $\tau$. Even accounting for the change in $\langle\tau\rangle$, this moves cannot lead to a short ranged $p(\tau)$ for $x$ if $p(\tau)$ of $z$ has a long tail (we will show it later in the end of this subsection). All in, we expect burstiness to be preserved moving up, and destroyed moving down in the hierarchy of levels.

As we said above, from Eq. (3.22), long-range correlations $\gamma > 1$ can be originated by to two possible sources: the tail of $p(\tau)$ (burstiness) and the tail of $C_\tau(k)$. The analysis above shows their different role at different levels in the hierarchy: $\gamma$ is preserved moving down, but there is a transfer of information from $p(\tau)$ to $C_\tau(k)$. This can be better understood considering the following simplified set-up: suppose that at a level we observe a sequence $x$ coming from a renewal process such that

$$p(\tau) \propto \tau^{-\mu} \quad \text{and} \quad C_\tau(k) = \delta(k) \tag{3.36}$$

with $2 < \mu < 3$. Now we can consider the behavior of $z$, obtained by adding to $x$ other independent sequences. The long $\tau$'s (a long sequence of zeros) in Eq. (3.36) will be divided in two long sequences introducing at the same time a cut-off $\tau_c$ in $p(\tau)$ and non trivial correlations $C_\tau(k) \neq 0$ for large $k$. In such a case, long-range correlations ($\gamma_Z = \max\{\gamma_X, \gamma_Y\} > 1$) is solely due to $C_\tau(k) \neq 0$. Burstiness affects only $\hat{\gamma}$ for times $t < \tau_c$. An analogous result is expected in the generic case of a starting sequence $x$ with broad tails in both $p(\tau)$ and $C_\tau(k)$.

Now we consider the simplified set-up exposed in Eq. (3.36): $z$ is a sequence coming from a renewal process such that $p(\tau) \propto \tau^{-\mu}$ and $C_\tau(k) = \delta(k)$. Once given a fixed $0 \leq \langle\zeta\rangle \leq 1$, if we consider the random subtraction $x = \zeta z$ where each $z_j = 1$ is set to $z_j = 0$ with probability $\langle\zeta\rangle$, then the inter-event times' distribution of this new process is

$$\tilde{p}(\tau) = (1 - \langle\zeta\rangle)\, p(\tau) + \sum_{k=1}^{\infty} \langle\zeta\rangle^k \sum_{t_1+t_2+\ldots+t_k=r} \prod_{j=1}^{k} p(t_j). \tag{3.37}$$

Asymptotically $\tilde{p}(\tau)$ is dominated by the long tails of $(1 - \langle\zeta\rangle)\, p(\tau)$. In fact, if $\tau$ is large enough, once fixed $\overline{k} > 0$ eventually diverging with $\tau \to \infty$ and divided accordingly the sum over $k$ in the second term of the right hand side, then the term corresponding to the sum $k > \overline{k}$ is dominated by $\zeta^{\overline{k}}$ and arbitrary small, while the remaining finite sum over $k \leq \overline{k}$ is controlled again by the tail of $p(\tau)$.

## 3.3 Confidence Interval for Long-Range Correlation

The finite time estimator of the long-range correlation $\hat{\gamma}$ will be computed fitting Eq. (3.6) for a range of $t$, $t \in [t_{s'}, t_s]$. Now we will analyze the procedure used to obtain values for $t_{s'}$ and $t_s$ proposed by E. G. Altmann, G. Cristadoro and M. Degli Esposti in [43].

As we already know, the distinction between long-range and short-range correlation needs a finite-time estimate $\hat{\gamma}$ of the asymptotic exponent $\gamma$ of the random walker associated to a binary sequence, that means estimating the $\sigma_X^2(t) \propto t^\gamma$ relation and it is therefore essential to estimate the upper limit in $t$, $t_s$, for which we have enough accuracy to obtain an acceptable estimate $\hat{\gamma}$. They adopted the following procedure to estimate $t_s$. Considering a alternative binary sequence with the same length $N$ and a series of ones randomly placed in the sequence. For this sequence we know that $\gamma = 1$. They then considered a sequence of times $t_i$ equally distributed in the logarithmic scale of $t$ (they considered $\dfrac{t_{i+1}}{t_i} = 1.2$, with $i$ integer and $t_0 = 1$) and estimated the local exponent as

$$\hat{\gamma}_{\text{local}}(t_i) = \frac{\log_{10} \Delta\sigma_X^2(t_{i+1}) - \log_{10} \Delta\sigma_X^2(t_i)}{\log_{10}(1.2)}$$

For small $t$, $\hat{\gamma}_{\text{local}} \approx 1$ but as $t$ became larger, statistical fluctuations increased due to the finiteness of $N$. So they chose $t_s$ as the smallest $t_i$ for which $\{\hat{\gamma}_{\text{local}}(t_{i+1}), \hat{\gamma}_{\text{local}}(t_{i+2}), \hat{\gamma}_{\text{local}}(t_{i+3})\}$ were all outside $[0.9, 1.1]$. They

also verified that $t_s$ scales linearly with $N$.

Based on these results, a good estimate of $t_s$ is $t_s = \frac{N}{100}$. This rule have been used in the estimate of $\hat{\gamma}$ for all the experiments that we will analyze in the following chapter. The $t_s$ is only the upper limit and the estimate $\hat{\gamma}$ is performed through a linear regression fit (using the program *Grace*) in the time interval $t_{s'} < t < t_s = \frac{N}{100}$, where $t_{s'} \approx \frac{t_s}{100}$.

This result can be better understood thanks to the following figure, caught from [43].

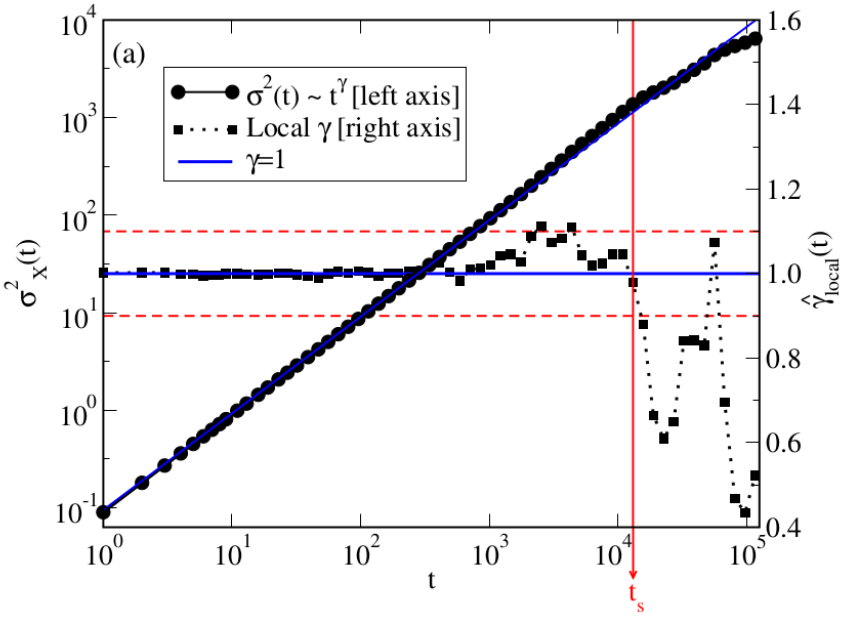



Figure 3.2: Determination of the time interval for the estimate of the long-range correlation exponent $\hat{\gamma}$. $\sigma_X^2(t)$ is shown as •for a random binary sequence of size $N = 10^6$ and 10% of ones. The local derivative is shown as ■ and agrees with the theoretical exponent $\gamma = 1$ until fluctuations starts for large $t$ (axis on the right). The time $t_s$ denotes the end of the interval of safe determination of $\gamma$, as explained above.

# Chapter 4

# Long-Range Correlations and Burstiness in different languages

Equipped with previous chapter's theoretical framework, here we will interpret observations in real texts, focusing on the comparison of our results between different languages and looking for differences and similarity in different translations.

## 4.1 Preliminary analysis

Before beginning our study on different languages we will do a preliminary analysis, using *War and Peace* by Leo Tolstoj in English, in which we will observe, thanks to real data, the behavior of $\sigma_X^2(t) \propto t^\gamma$ and we will propose a measure for the burstiness, $\frac{\sigma_\tau}{\langle\tau\rangle}$.

First of all we introduce the measure for the burstiness. In line with what we have studied in the previous chapter, that is that burstiness is measured as the broad tail of the distribution of inter-event times $p(\tau)$ (divergent $\sigma_\tau$), using the proposal of K. I. Goh and A. L. Barabasi in [36], we can immediatly

reach our measure. In fact K. I. Goh and A. L. Barabasi proposed as a measure for the burstiness the following:

$$B \equiv \frac{\frac{\sigma_\tau}{\langle\tau\rangle} - 1}{\frac{\sigma_\tau}{\langle\tau\rangle} + 1} = \frac{\sigma_\tau - \langle\tau\rangle}{\sigma_\tau + \langle\tau\rangle} = 1 - 2\frac{\langle\tau\rangle}{\sigma_\tau + \langle\tau\rangle} =$$

$$= 1 - 2\left(\frac{\sigma_\tau + \langle\tau\rangle}{\langle\tau\rangle}\right)^{-1} = 1 - 2\left(1 + \frac{\sigma_\tau}{\langle\tau\rangle}\right)^{-1} \tag{4.1}$$

Hence we can consider $B' = \dfrac{\sigma_\tau}{\langle\tau\rangle}$ as a measure for the burstiness.
It's easily observable that our burstiness measure for a Poisson process is equal to 1.

Now we can study and observe long-range correlations and burstiness on real data. First of all in the following figures there are linear and $\log - \log$ plots of $\sigma_X^2(t)$ for the space " ", the symbol *"e"* and the word " *prince* ".

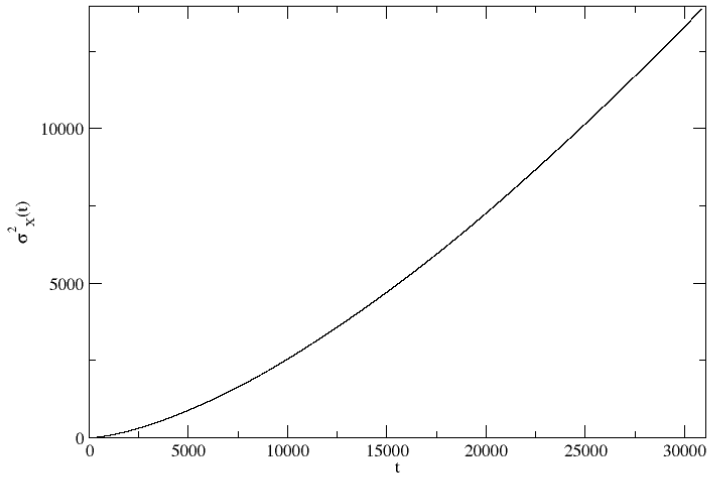



Figure 4.1: $\sigma^2_X(t)$ plot for the space " "

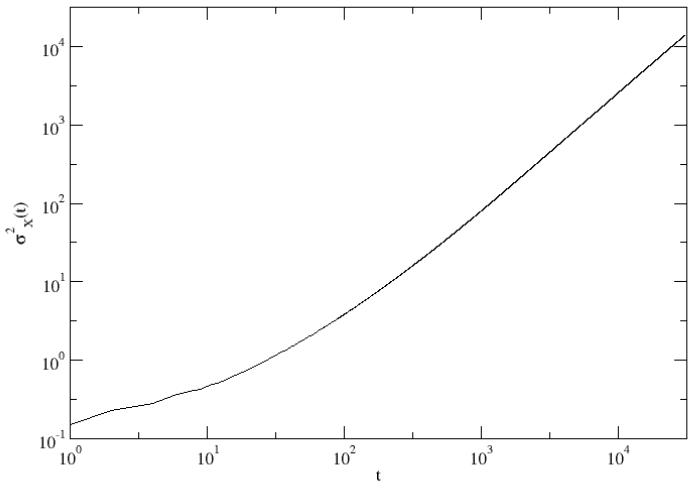



Figure 4.2: $\sigma^2_X(t)$ $\log - \log$ plot for the space " "

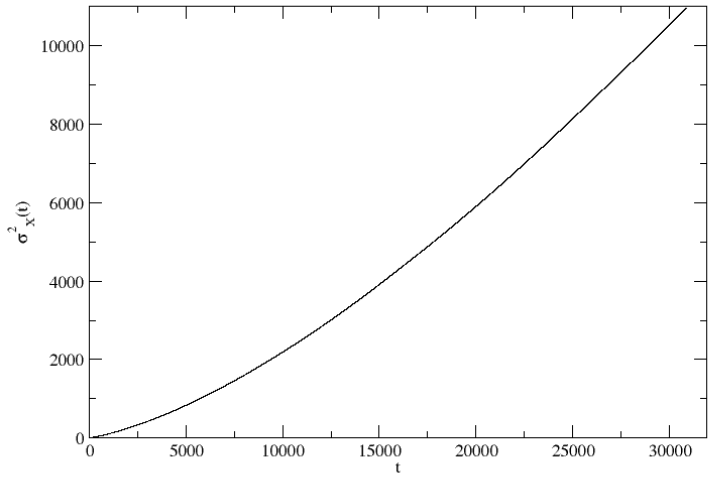



Figure 4.3: $\sigma_X^2(t)$ plot for the symbol *"e"*

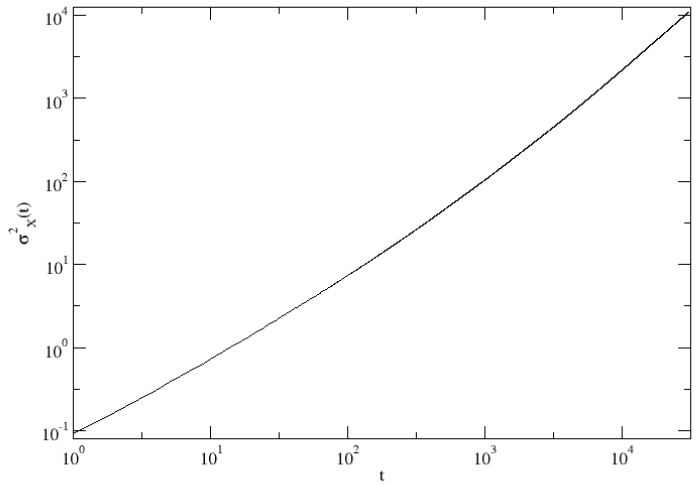



Figure 4.4: $\sigma_X^2(t)$ $\log - \log$ plot for the symbol *"e"*

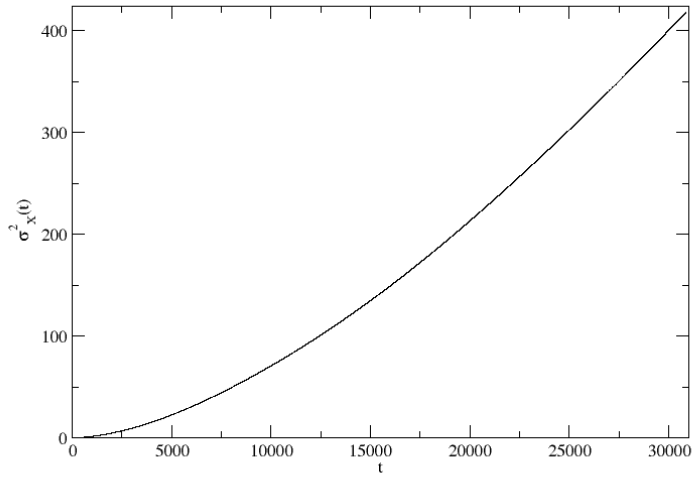



Figure 4.5: $\sigma^2_X(t)$ plot for the word " *prince* "

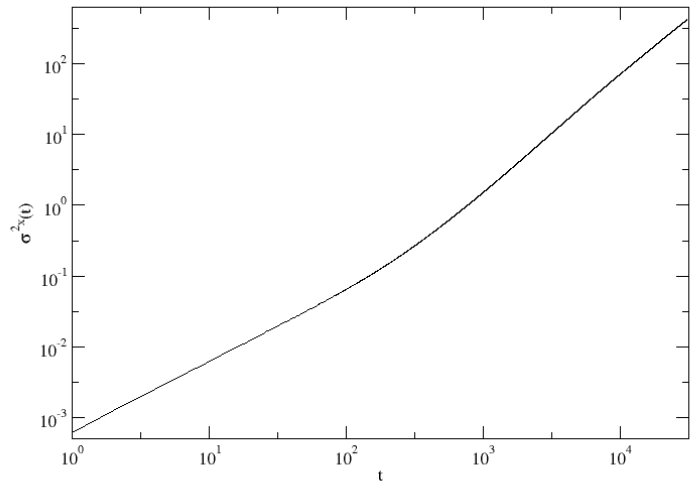



Figure 4.6: $\sigma^2_X(t)$ $\log - \log$ plot for the word " *prince* "

As it can be easily seen in the previous figures, especially in the $\log - \log$ plots, $\sigma_X^2(t)$ follows Eq. (3.32) in all these 3 cases, in fact in the first part of the plot, $\sigma_X^2(t)$ isn't dominated by an exponential increase, but after a certain time it's clear that it follows his asymptotical behavior described by $\sigma_X^2(t) \propto t^\gamma$. In the following figures there are the same $\log - \log$ plots shown before with relative approximation in the range described in Section **3.3**.

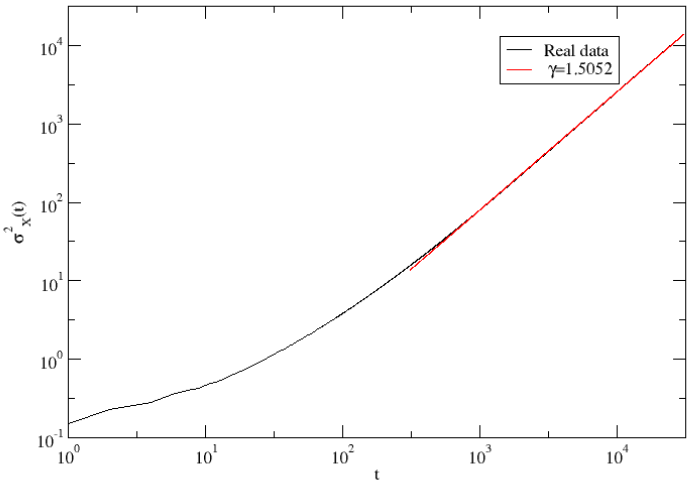



Figure 4.7: $\sigma_X^2(t)$ $\log - \log$ plot for the space " ", $\gamma_X = 1.5052 \pm 0.611149 \times 10^{-4}$

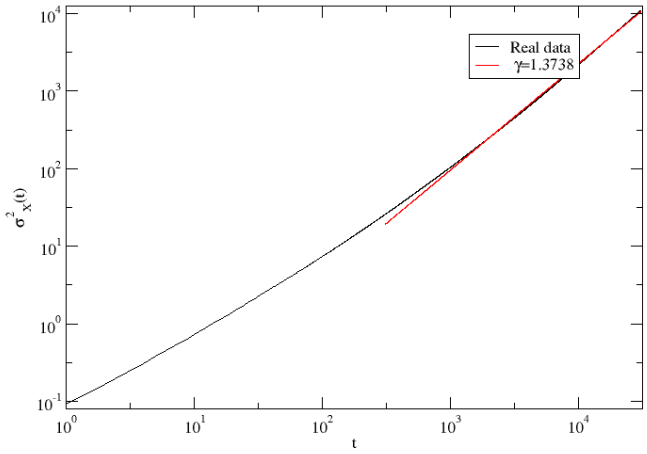



Figure 4.8: $\sigma_X^2(t)$ log − log plot for the symbol *"e"*, $\gamma_X = 1.3738 \pm 0.24585 \times 10^{-3}$

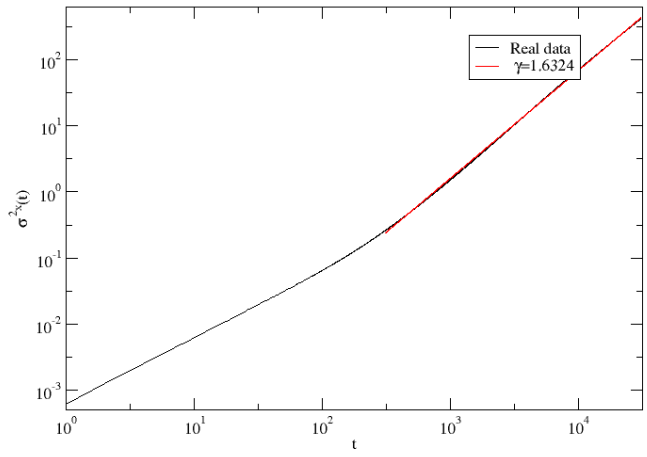



Figure 4.9: $\sigma_X^2(t)$ log − log plot for the word *" prince "*, $\gamma_X = 1.6324 \pm 0.14617 \times 10^{-3}$

Similarly we can study our burstiness measure for these words. First of all, in the following figures there are their relative plots of distribution of $P(\tau)$.

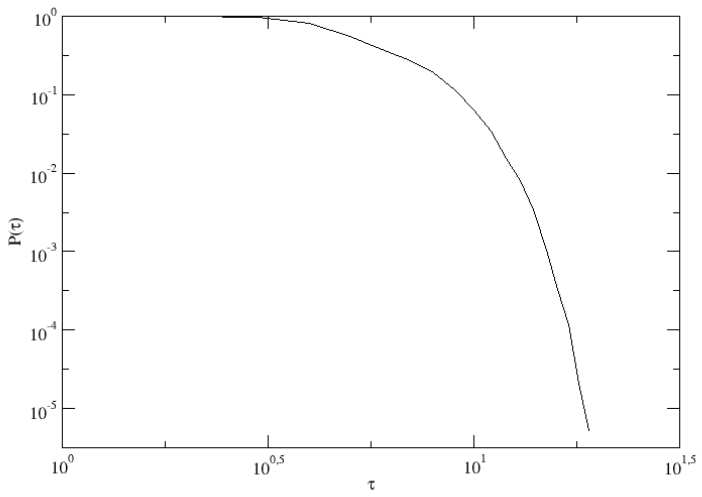



Figure 4.10: $P(\tau)$ $\log - \log$ plot for the space " "

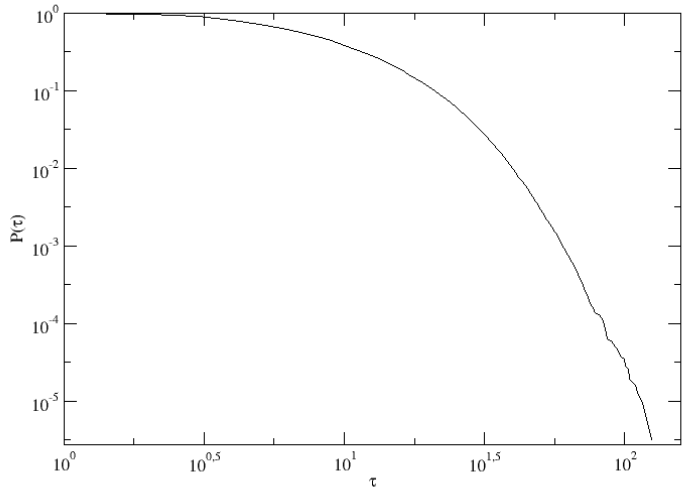



Figure 4.11: $P(\tau)$ $\log - \log$ plot for the symbol *"e"*

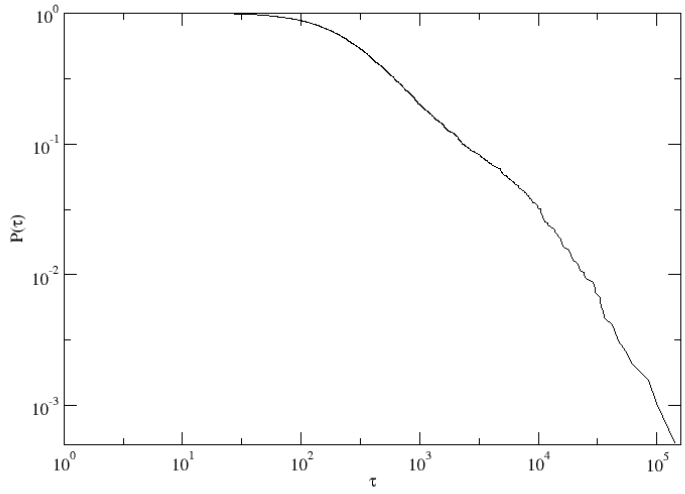



Figure 4.12: $P(\tau)$ $\log - \log$ plot for the word *" prince "*

From these plots, and even from our experience, we may expect that *" "* and *"e"* should have a small burstiness coefficient (we expect that they are almost equally distributed all along the text), while the word *" prince "*, that is much more present in particular contexts, should have a bigger burstiness coefficient. And this is exactly what happens, in fact $B'(\text{" "}) = 0.52744$, $B'(\text{"e"}) = 0.92777$ and $B'(\text{" prince "}) = 3.9227$.

## 4.2 Distinct analysis on languages

Now that we have observed on real data what we had up to now only theorically studied, we can go on with our analysis and focus our attention on analogies and differences of burstiness and long-range correlations in different languages.

First of all we will analyze, language per language, long-range correlation and burstiness behaviors for various condition $\alpha$ in different translations of the same book, *War and Peace* by Leo Tolstoj. For this analysis we will use the procedure used and exposed by E. G. Altmann, G. Cristadoro and M. Degli Esposti in [43].

For each language 43 binary sequences will be analyzed separately: vowels and consonants, 20 at the letter level (blank space and the 19 most frequent characters), and 21 at the word level (7 most frequent words, 7 most frequent nouns, and 7 words with frequency matched to the frequency of the nouns). In the following tabs there are all the results, obtained thanks to my experiments. In particular, for each condition $\alpha$ (and consequently each binary sequence), are shown number of occurrences ($f$), burstiness measure ($B'$), long-range correlation exponent evaluate ($\hat{\gamma}$) and standard deviation of $\hat{\gamma}$ ($\sigma_{\hat{\gamma}}$).

Moreover, as we said in the previous chapter, we are interested in searching the origin of long-range correlations, so in distinguishing if the obtained value of $\hat{\gamma}$ is due to burstiness corresponding to $p(\tau)$ with diverging $\sigma_\tau$ or diverging $\sum C_\tau(k)$. In order to be able to distinguish between these two possible origins we will compare asymptotic behavior of $x$ with the asymptotic behavior of two fictitious sequences $x_1$ and $x_2$ obtained from $x$ in the following ways:

- $x_1$ is obtained shuffling the sequence of $\{0, 1\}$ and this particular shuffle destroys every kind of correlations;
- $x_2$ is obtained shuffling the sequence of inter-event times $\tau_i$ and this particular shuffle destroys correlations due to $\sum C_\tau(k)$ and preserves correlations due to $p(\tau)$.

Using this result, in the following tab there are, for all sequences obtained at letter levels, estimates of $\gamma_1$ ($\hat{\gamma_1}$) and of $\gamma_2$ ($\hat{\gamma_2}$) too, where $\gamma_i$ is the exponent of $\sigma^2_{X_i}(t) \propto t^{\gamma_i}$.

***War and Peace* in English,** $N = 3086648$

| $\alpha$ | $f$ | $B'$ | $\hat{\gamma}$ | $\sigma_{\hat{\gamma}} \times 10^3$ | $\hat{\gamma_1}$ | $\hat{\gamma_2}$ |
|---|---|---|---|---|---|---|
| vowel | 946517 | 0.61797 | 1.5271 | 0.20860 | 1.0072 | 1.007 |
| consonant | 1566573 | 0.83332 | 1.445 | 0.11434 | 1.0552 | 1.0461 |
| " " | 572625 | 0.52981 | 1.5052 | 0.061149 | 1.0036 | 0.9785 |
| "e" | 313039 | 0.92834 | 1.3738 | 0.24585 | 0.97484 | 1.0211 |
| "t" | 224512 | 0.95566 | 1.3767 | 0.14824 | 1.0223 | 1.063 |
| "a" | 204424 | 0.93075 | 1.3996 | 0.20916 | 0.9293 | 1.009 |
| "o" | 191494 | 0.96383 | 1.4392 | 0.25690 | 1.0109 | 0.98787 |
| "n" | 183129 | 0.91489 | 1.2393 | 0.21009 | 0.98852 | 1.0094 |
| "i" | 172641 | 0.94681 | 1.4624 | 0.25619 | 1.01 | 0.96923 |
| "h" | 166520 | 0.90050 | 1.4679 | 0.16054 | 1.0471 | 0.98485 |
| "s" | 162128 | 1.0089 | 1.3043 | 0.16479 | 1.007 | 0.95442 |
| "r" | 146890 | 0.96476 | 1.3514 | 0.079064 | 1.0301 | 1.0079 |
| "d" | 117753 | 0.95994 | 1.4482 | 0.17842 | 1.003 | 0.96837 |
| "l" | 96037 | 1.0988 | 1.2278 | 0.11851 | 1.0086 | 0.90691 |
| "u" | 64919 | 0.99276 | 1.2273 | 0.052257 | 0.98201 | 0.99009 |
| "m" | 61283 | 1.0386 | 1.2674 | 0.094954 | 0.95521 | 1.0114 |
| "c" | 60659 | 1.0290 | 1.512 | 0.20701 | 1.0679 | 1.0227 |
| "w" | 58930 | 0.99698 | 1.2582 | 0.14607 | 0.94662 | 1.025 |
| "f" | 54507 | 1.0829 | 1.4713 | 0.22212 | 1.0368 | 1.0302 |
| "g" | 50909 | 1.0314 | 1.4749 | 0.24452 | 1.0574 | 0.98835 |
| "y" | 45936 | 1.0513 | 1.3307 | 0.056915 | 1.021 | 1.0044 |
| "p" | 44717 | 1.0965 | 1.4642 | 0.23978 | 1.0058 | 0.96415 |

***War and Peace* in English,** $N = 3086648$

| $\alpha$ | $f$ | $B'$ | $\hat{\gamma}$ | $\sigma_{\hat{\gamma}} \times 10^3$ | $\hat{\gamma_1}$ | $\hat{\gamma_2}$ |
|---|---|---|---|---|---|---|
| ” pierre ” | 1963 | 6.8589 | 1.7244 | 0.22683 | 0.94582 | 1.642 |
| ” prince ” | 1928 | 3.9026 | 1.6324 | 0.14617 | 0.98025 | 1.4101 |
| ” so ” | 1902 | 1.1816 | 1.1297 | 0.14574 | 0.95724 | 0.98121 |
| ” an ” | 1628 | 1.1329 | 1.1629 | 0.18907 | 0.97498 | 1.0621 |
| ” natasha ” | 1213 | 6.0441 | 1.6854 | 0.15081 | 0.96974 | 1.6652 |
| ” man ” | 1189 | 1.4277 | 1.3983 | 0.16457 | 0.89977 | 1.0782 |
| ” t ” | 1159 | 1.9598 | 1.3655 | 0.11500 | 0.94903 | 1.1954 |
| ” andrew ” | 1143 | 4.1074 | 1.6555 | 0.19702 | 1.0186 | 1.4387 |
| ” could ” | 1115 | 1.2850 | 1.1107 | 0.15293 | 0.97007 | 1.0859 |
| ” we ” | 1069 | 1.9264 | 1.3716 | 0.30257 | 0.98505 | 1.1063 |
| ” time ” | 929 | 1.1027 | 1.1091 | 0.082071 | 1.0381 | 1.0204 |
| ” princess ” | 916 | 5.4370 | 1.6668 | 0.16860 | 1.034 | 1.6061 |
| ” face ” | 893 | 1.4457 | 1.2249 | 0.059158 | 0.96738 | 1.1763 |
| ” french ” | 881 | 2.2884 | 1.5068 | 0.17895 | 1.0196 | 1.2611 |
| ” the ” | 34545 | 1.1409 | 1.5647 | 0.12969 | 0.96329 | 1.0262 |
| ” and ” | 22227 | 0.8813 | 1.1965 | 0.064044 | 1.0127 | 0.93914 |
| ” to ” | 16675 | 1.0616 | 1.2398 | 0.25609 | 0.98013 | 0.96386 |
| ” of ” | 14889 | 1.1752 | 1.5587 | 0.20810 | 0.96831 | 1.0085 |
| ” a ” | 10551 | 1.1230 | 1.1752 | 0.17585 | 0.98497 | 0.99979 |
| ” he ” | 10002 | 1.9056 | 1.381 | 0.17895 | 0.98183 | 1.1786 |
| ” in ” | 8979 | 1.0436 | 1.1471 | 0.071755 | 1.0164 | 0.95459 |

***War and Peace* in French, $N = 2789763$**

| $\alpha$ | $f$ | $B'$ | $\hat{\gamma}$ | $\sigma_{\hat{\gamma}} \times 10^3$ | $\hat{\gamma_1}$ | $\hat{\gamma_2}$ |
|---|---|---|---|---|---|---|
| vowel | 997609 | 0.67025 | 1.321 | 0.18528 | 1.0165 | 0.97224 |
| consonant | 1216851 | 0.79767 | 1.3792 | 0.24494 | 0.99687 | 1.0209 |
| " " | 505476 | 0.61263 | 1.3967 | 0.052013 | 1.016 | 0.97946 |
| "e" | 378513 | 0.86628 | 1.3551 | 0.23265 | 0.94734 | 1.0257 |
| "a" | 197381 | 0.92865 | 1.3593 | 0.15307 | 0.99429 | 0.96052 |
| "s" | 174233 | 1.1071 | 1.2519 | 0.088000 | 1.0091 | 1.0023 |
| "i" | 169672 | 0.92964 | 1.2605 | 0.15523 | 1.0417 | 0.99406 |
| "t" | 165898 | 0.93999 | 1.2338 | 0.22159 | 0.99575 | 1.0594 |
| "n" | 157399 | 0.96502 | 1.1352 | 0.15076 | 0.97408 | 1.0377 |
| "r" | 148871 | 0.94514 | 1.3058 | 0.23110 | 1.0076 | 1.0032 |
| "u" | 133638 | 0.94943 | 1.2081 | 0.11381 | 1.0604 | 0.87378 |
| "l" | 125772 | 0.97877 | 1.2968 | 0.12287 | 1.0359 | 0.97792 |
| "o" | 118405 | 0.93921 | 1.3074 | 0.083222 | 1.0093 | 1.0219 |
| "d" | 77194 | 0.94100 | 1.3535 | 0.37567 | 0.95791 | 1.0294 |
| "c" | 67807 | 0.96454 | 1.2191 | 0.16425 | 1.0287 | 0.97332 |
| "" | 63908 | 1.0309 | 1.2748 | 0.19256 | 0.96787 | 0.93485 |
| "p" | 61729 | 0.98386 | 1.3276 | 0.27603 | 0.9946 | 0.96828 |
| "m" | 59982 | 1.0788 | 1.2492 | 0.12959 | 0.99511 | 0.95589 |
| "v" | 38609 | 0.98788 | 1.2292 | 0.15740 | 0.91457 | 1.0258 |
| "q" | 23732 | 0.98872 | 1.2322 | 0.19524 | 0.93112 | 0.91942 |
| "f" | 23265 | 1.0956 | 1.2479 | 0.19474 | 0.9696 | 0.99861 |
| "h" | 20041 | 1.0290 | 1.3492 | 0.31584 | 0.96458 | 0.95534 |

***War and Peace* in French,** $N = 2789763$

| $\alpha$ | $f$ | $B'$ | $\hat{\gamma}$ | $\sigma_{\hat{\gamma}} \times 10^3$ | $\hat{\gamma_1}$ | $\hat{\gamma_2}$ |
|---|---|---|---|---|---|---|
| ” prince ” | 1307 | 3.2224 | 1.5819 | 0.15491 | 1.0541 | 1.4204 |
| ” meme ” | 1301 | 1.0409 | 1.0614 | 0.043460 | 0.99424 | 0.99619 |
| ” pierre ” | 1252 | 4.4686 | 1.6915 | 0.23414 | 0.98082 | 1.5784 |
| ” nous ” | 1156 | 1.9510 | 1.2138 | 0.27870 | 1.0123 | 1.105 |
| ” natacha ” | 811 | 5.0081 | 1.6755 | 0.17735 | 0.96522 | 1.6194 |
| ” tous ” | 806 | 1.1246 | 1.0814 | 0.10447 | 0.99956 | 1.0277 |
| ” yeux ” | 754 | 1.2288 | 1.119 | 0.11645 | 0.9741 | 1.118 |
| ” j ” | 750 | 1.5411 | 1.1634 | 0.11118 | 0.94583 | 1.1325 |
| ” andre’ ” | 731 | 3.2587 | 1.6163 | 0.24708 | 1.021 | 1.4678 |
| ” ai ” | 716 | 1.4780 | 1.2062 | 0.11077 | 1.0361 | 1.1076 |
| ” rostow ” | 635 | 3.6402 | 1.6197 | 0.24562 | 0.93567 | 1.4483 |
| ” e’te’ ” | 621 | 1.1495 | 1.0427 | 0.092373 | 0.97583 | 1.0067 |
| ” princesse ” | 603 | 4.7442 | 1.6425 | 0.18251 | 1.0299 | 1.591 |
| ” fait ” | 602 | 1.1014 | 1.0812 | 0.10262 | 1.0197 | 1.0633 |
| ” de ” | 21367 | 1.0004 | 1.2472 | 0.16321 | 1.0318 | 0.96458 |
| ” et ” | 15457 | 0.82029 | 1.1889 | 0.22260 | 1.013 | 1.018 |
| ” la ” | 13070 | 1.0582 | 1.215 | 0.069586 | 1.0332 | 1.0272 |
| ” a’ ” | 11825 | 0.99179 | 1.1259 | 0.16958 | 0.97377 | 0.99712 |
| ” le ” | 11375 | 1.0675 | 1.2878 | 0.22378 | 1.0308 | 1.0207 |
| ” il ” | 10377 | 1.3124 | 1.3239 | 0.22268 | 0.99126 | 1.0498 |
| ” l ” | 9142 | 1.0755 | 1.3439 | 0.27476 | 1.0034 | 1.0404 |

***War and Peace* in German,** $N = 3602335$

| $\alpha$ | $f$ | $B'$ | $\hat{\gamma}$ | $\sigma_{\hat{\gamma}} \times 10^3$ | $\hat{\gamma_1}$ | $\hat{\gamma_2}$ |
|---|---|---|---|---|---|---|
| vowel | 1149355 | 0.62397 | 1.3008 | 0.20550 | 0.99585 | 1.0509 |
| consonant | 1867357 | 0.89555 | 1.3805 | 0.15053 | 0.97903 | 1.033 |
| " " | 582729 | 0.54066 | 1.5489 | 0.097998 | 0.9806 | 0.99867 |
| "e" | 502796 | 0.86223 | 1.5034 | 0.14398 | 1.0272 | 1.0129 |
| "n" | 309851 | 0.95298 | 1.2873 | 0.11169 | 1.0446 | 0.95613 |
| "i" | 230763 | 0.88366 | 1.4988 | 0.20140 | 0.99224 | 0.96838 |
| "r" | 214406 | 0.91988 | 1.4495 | 0.23139 | 0.97666 | 1.0106 |
| "s" | 209920 | 1.0148 | 1.3181 | 0.11601 | 0.99819 | 1.0518 |
| "a" | 198389 | 0.92288 | 1.5074 | 0.21003 | 0.96507 | 0.9691 |
| "t" | 171030 | 0.99021 | 1.1882 | 0.083631 | 0.99541 | 0.9404 |
| "h" | 154327 | 0.93429 | 1.4256 | 0.21660 | 0.99746 | 0.98877 |
| "d" | 148939 | 0.89580 | 1.4043 | 0.12591 | 1.0662 | 0.99609 |
| "u" | 136121 | 0.95034 | 1.2772 | 0.15781 | 0.94942 | 1.0204 |
| "l" | 102135 | 1.0801 | 1.2533 | 0.059517 | 0.98211 | 0.98513 |
| "c" | 95667 | 0.96185 | 1.3419 | 0.11811 | 1.0059 | 0.95614 |
| "g" | 86343 | 0.99539 | 1.2069 | 0.10606 | 0.9223 | 0.94324 |
| "o" | 81286 | 1.0334 | 1.5349 | 0.14282 | 1.0198 | 1.0404 |
| "m" | 80323 | 1.1216 | 1.299 | 0.19548 | 0.95629 | 0.98822 |
| "b" | 56096 | 0.98659 | 1.2601 | 0.095810 | 0.97285 | 0.99889 |
| "w" | 54645 | 1.0169 | 1.3357 | 0.15527 | 0.99574 | 1.0031 |
| "f" | 51746 | 1.0485 | 1.3739 | 0.16503 | 1.057 | 0.98509 |
| "z" | 38314 | 1.0052 | 1.1843 | 0.083416 | 1.0233 | 0.99831 |

***War and Peace* in German,** $N = 3602335$

| $\alpha$ | $f$ | $B'$ | $\hat{\gamma}$ | $\sigma_{\hat{\gamma}} \times 10^3$ | $\hat{\gamma_1}$ | $\hat{\gamma_2}$ |
|---|---|---|---|---|---|---|
| ” pierre ” | 1947 | 6.7635 | 1.7192 | 0.23027 | 0.95166 | 1.626 |
| ” doch ” | 1942 | 1.2176 | 1.2069 | 0.19368 | 1.0197 | 1.0668 |
| ” furst ” | 1409 | 4.4620 | 1.636 | 0.12477 | 1.0359 | 1.5006 |
| ” alle ” | 1409 | 1.2112 | 1.1196 | 0.048206 | 0.97702 | 1.0406 |
| ” natascha ” | 1157 | 6.4678 | 1.6873 | 0.12079 | 1.0072 | 1.6031 |
| ” durch ” | 1153 | 1.2041 | 1.1668 | 0.1482 | 0.93579 | 0.98965 |
| ” andrej ” | 1138 | 4.0486 | 1.662 | 0.20274 | 1.0073 | 1.4835 |
| ” jetzt ” | 1132 | 1.1804 | 1.0651 | 0.16538 | 1.0263 | 1.0238 |
| ” augen ” | 878 | 1.3574 | 1.1848 | 0.065447 | 0.99162 | 1.1396 |
| ” gesicht ” | 865 | 1.5477 | 1.2079 | 0.10690 | 0.90117 | 1.175 |
| ” mich ” | 842 | 1.6716 | 1.3117 | 0.088909 | 0.92484 | 1.1625 |
| ” prinzessin ” | 833 | 5.4406 | 1.6588 | 0.18420 | 1.0121 | 1.608 |
| ” oder ” | 880 | 1.2541 | 1.1243 | 0.15328 | 1.0956 | 1.0106 |
| ” konnte ” | 810 | 1.1629 | 1.1383 | 0.069934 | 0.97847 | 1.0497 |
| ” und ” | 21990 | 0.86393 | 1.2799 | 0.18450 | 0.97306 | 1.0674 |
| ” die ” | 15265 | 1.1412 | 1.3322 | 0.070994 | 0.97738 | 0.99942 |
| ” der ” | 13509 | 1.1870 | 1.4671 | 0.22934 | 1.0352 | 1.0736 |
| ” er ” | 10625 | 1.7130 | 1.3459 | 0.21846 | 0.9521 | 1.1623 |
| ” sie ” | 9171 | 1.8521 | 1.5723 | 0.066114 | 1.0162 | 1.0998 |
| ” zu ” | 8757 | 1.0807 | 1.1751 | 0.11354 | 1.038 | 1.0305 |
| ” in ” | 7789 | 1.0558 | 1.1643 | 0.10103 | 1.026 | 0.95713 |

***War and Peace* in Italian, $N = 3458573$**

| $\alpha$ | $f$ | $B'$ | $\hat{\gamma}$ | $\sigma_{\hat{\gamma}} \times 10^3$ | $\hat{\gamma_1}$ | $\hat{\gamma_2}$ |
|---|---|---|---|---|---|---|
| vowel | 1313825 | 0.53241 | 1.4171 | 0.072899 | 0.99932 | 0.94296 |
| consonant | 1521057 | 0.85645 | 1.4055 | 0.11720 | 0.92064 | 1.0171 |
| " " | 583357 | 0.58245 | 1.4059 | 0.080931 | 1.004 | 1.0064 |
| "e" | 333047 | 0.97890 | 1.4982 | 0.17171 | 0.97377 | 1.0564 |
| "a" | 328215 | 1.0211 | 1.5062 | 0.15785 | 0.98632 | 1.0016 |
| "i" | 289985 | 0.98436 | 1.3361 | 0.094486 | 0.95842 | 0.98606 |
| "o" | 270033 | 0.95758 | 1.405 | 0.13287 | 0.99126 | 1.0065 |
| "n" | 199391 | 0.97338 | 1.2253 | 0.063119 | 1.0181 | 1.0124 |
| "r" | 181154 | 0.93509 | 1.4043 | 0.26753 | 0.97877 | 0.99478 |
| "l" | 172686 | 1.0669 | 1.2655 | 0.17196 | 1.0013 | 1.079 |
| "t" | 168729 | 1.0975 | 1.3018 | 0.13559 | 1.0038 | 0.9759 |
| "s" | 167062 | 1.0820 | 1.2436 | 0.088212 | 1.0304 | 0.98825 |
| "c" | 130277 | 1.0220 | 1.3373 | 0.15669 | 1.0111 | 0.95912 |
| "d" | 102241 | 0.95956 | 1.3033 | 0.24882 | 0.96055 | 1.0119 |
| "u" | 92545 | 0.97957 | 1.3541 | 0.077144 | 0.99887 | 1.1062 |
| "p" | 84292 | 1.0418 | 1.4631 | 0.23805 | 1.0193 | 0.99873 |
| "v" | 70334 | 1.0956 | 1.4661 | 0.15796 | 0.89571 | 1.0174 |
| "m" | 69601 | 1.0369 | 1.3566 | 0.14720 | 0.97397 | 1.0028 |
| "g" | 48089 | 1.0981 | 1.3201 | 0.087395 | 0.9766 | 0.95999 |
| "" | 36345 | 1.1663 | 1.3601 | 0.071071 | 1.0473 | 1.0105 |
| "h" | 30204 | 1.0443 | 1.2992 | 0.16014 | 1.026 | 0.97679 |
| "f" | 27381 | 1.1530 | 1.3756 | 0.32566 | 0.96241 | 0.95043 |

***War and Peace* in Italian,** $N = 3458573$

| $\alpha$ | $f$ | $B'$ | $\hat{\gamma}$ | $\sigma_{\hat{\gamma}} \times 10^3$ | $\hat{\gamma_1}$ | $\hat{\gamma_2}$ |
|---|---|---|---|---|---|---|
| ” pierre ” | 2013 | 6.8904 | 1.7295 | 0.23561 | 0.92572 | 1.6456 |
| ” se ” | 2009 | 1.1407 | 1.0945 | 0.041174 | 1.0566 | 0.97375 |
| ” principe ” | 1935 | 3.8833 | 1.6309 | 0.15954 | 1.0039 | 1.4323 |
| ” alla ” | 1888 | 1.1340 | 1.0874 | 0.17638 | 1.0405 | 0.96481 |
| ” natascia ” | 1238 | 6.1838 | 1.6927 | 0.15411 | 1.022 | 1.6108 |
| ” cosa ” | 1228 | 1.1998 | 1.1941 | 0.16642 | 1.0037 | 1.0349 |
| ” andre’j ” | 1076 | 3.9794 | 1.661 | 0.21786 | 0.99681 | 1.4637 |
| ” delle ” | 1072 | 1.2827 | 1.2328 | 0.13926 | 1.0183 | 1.1139 |
| ” rosto’v ” | 931 | 4.4101 | 1.6837 | 0.18658 | 1.0173 | 1.4825 |
| ” questo ” | 902 | 1.2867 | 1.1902 | 0.20203 | 0.95068 | 1.0585 |
| ” occhi ” | 839 | 1.6834 | 1.2171 | 0.097512 | 0.99379 | 1.1707 |
| ” ha ” | 837 | 1.5979 | 1.1576 | 0.096694 | 1.0411 | 1.1171 |
| ” viso ” | 749 | 1.4485 | 1.2154 | 0.083104 | 0.99308 | 1.1476 |
| ” principessina ” | 742 | 5.2575 | 1.6537 | 0.20808 | 0.96939 | 1.5782 |
| ” e ” | 20515 | 0.94976 | 1.2496 | 0.14829 | 0.97573 | 0.97503 |
| ” di ” | 18859 | 1.1035 | 1.2559 | 0.25132 | 1.0016 | 0.97234 |
| ” che ” | 14586 | 1.0660 | 1.1966 | 0.15625 | 1.0319 | 1.009 |
| ” il ” | 12801 | 1.1062 | 1.3223 | 0.069590 | 0.96877 | 1.0458 |
| ” la ” | 12440 | 1.1171 | 1.392 | 0.10574 | 0.96935 | 0.99463 |
| ” a ” | 10152 | 1.0839 | 1.1625 | 0.15730 | 0.99241 | 1.0877 |
| ” si ” | 8776 | 1.1225 | 1.268 | 0.13450 | 0.98217 | 1.0381 |

These results are better observable in the following figures, where, for each language, there are four graphs for different sequences of the quantities $B' = \dfrac{\sigma_\tau}{\langle\tau\rangle}$ and $\hat{\gamma}$.

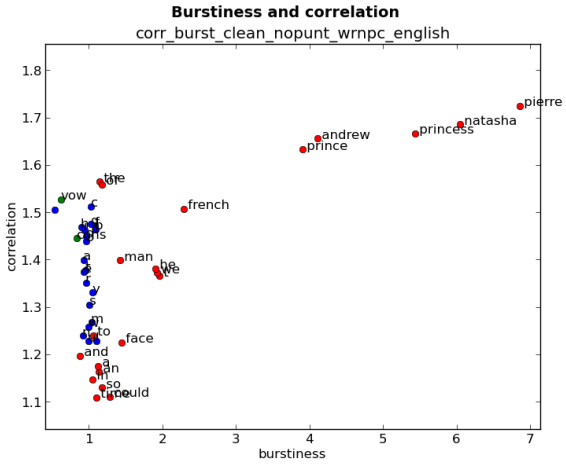



Figure 4.13: Burstiness-Correlation diagram for all 43 binary sequences studied in *War and Peace* in English. Green points are vowels (vow) and consonants (cons), blue points are symbols and red points are words.

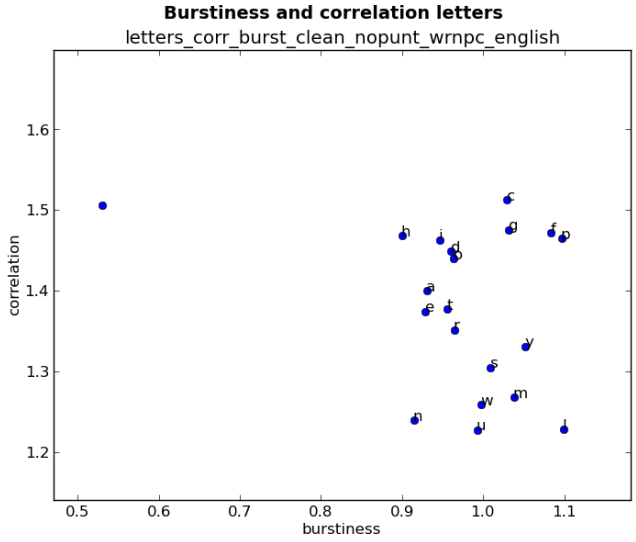



Figure 4.14: Burstiness-Correlation diagram for all symbols studied in *War and Peace* in English.

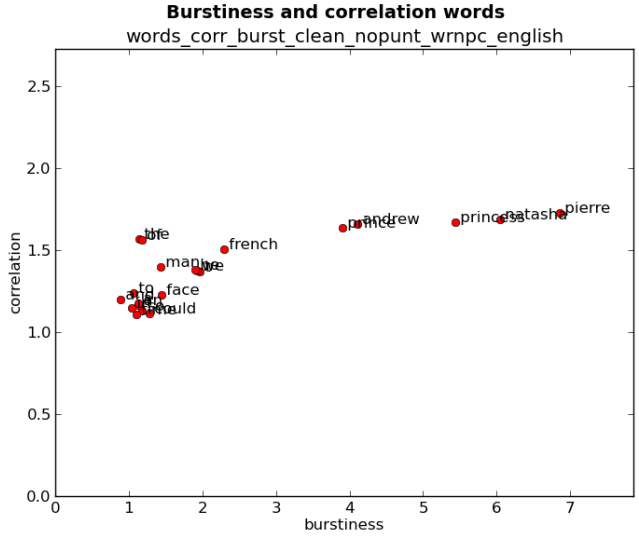



Figure 4.15: Burstiness-Correlation diagram for all words studied in *War and Peace* in English.

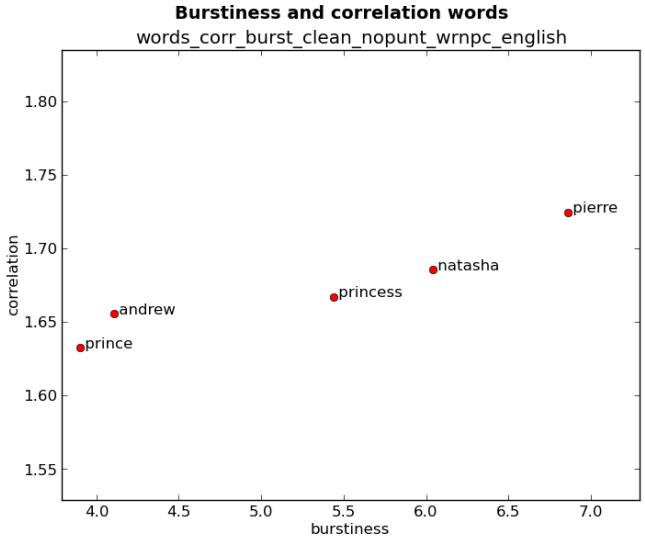



Figure 4.16: Burstiness-Correlation diagram for those words studied in *War and Peace* in English with high values of $B'$ and $\hat{\gamma}$.

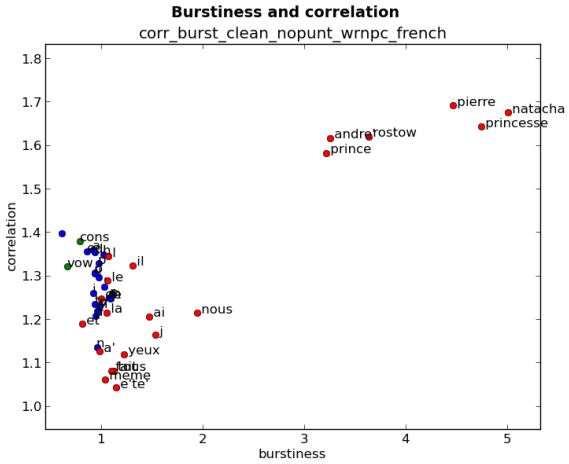



Figure 4.17: Burstiness-Correlation diagram for all 43 binary sequences studied in *War and Peace* in French. Green points are vowels (vow) and consonants (cons), blue points are symbols and red points are words.

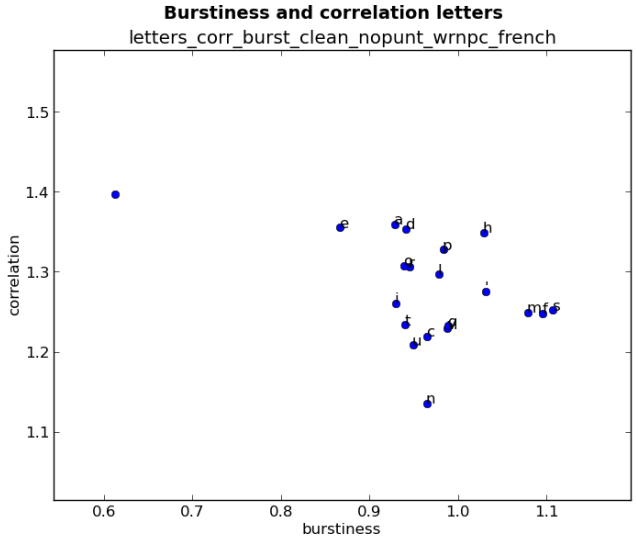



Figure 4.18: Burstiness-Correlation diagram for all symbols studied in *War and Peace* in French.

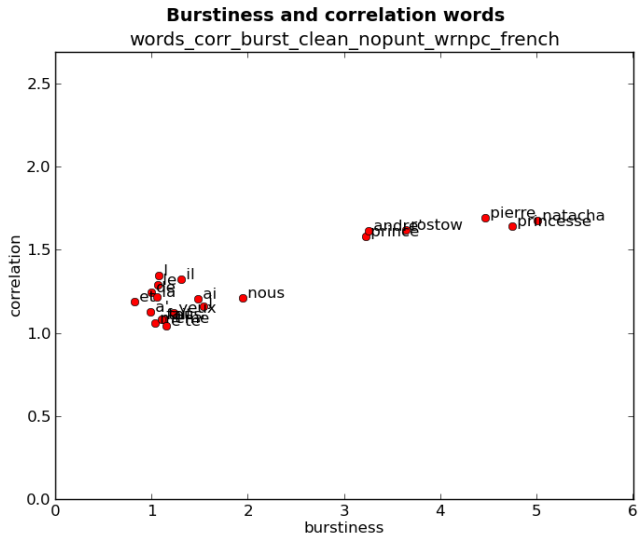



Figure 4.19: Burstiness-Correlation diagram for all words studied in *War and Peace* in French.

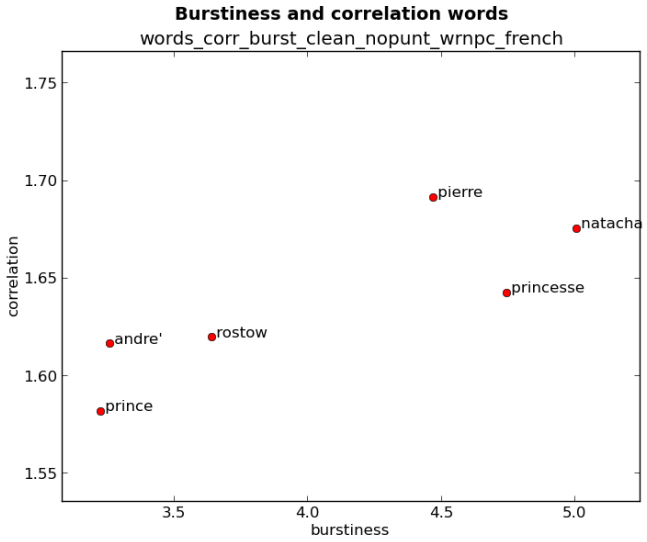



Figure 4.20: Burstiness-Correlation diagram for those words studied in *War and Peace* in French with high values of $B'$ and $\hat{\gamma}$.

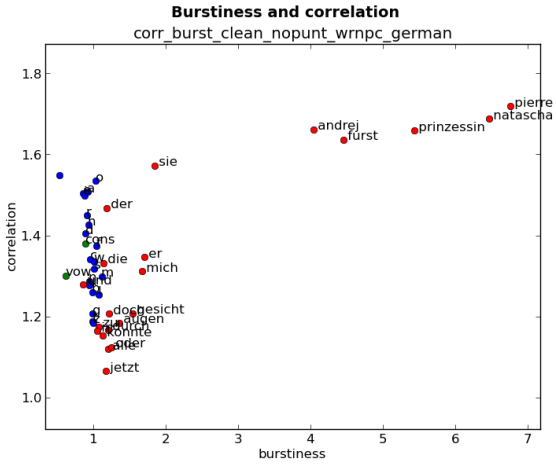



Figure 4.21: Burstiness-Correlation diagram for all 43 binary sequences studied in *War and Peace* in German. Green points are vowels (vow) and consonants (cons), blue points are symbols and red points are words.

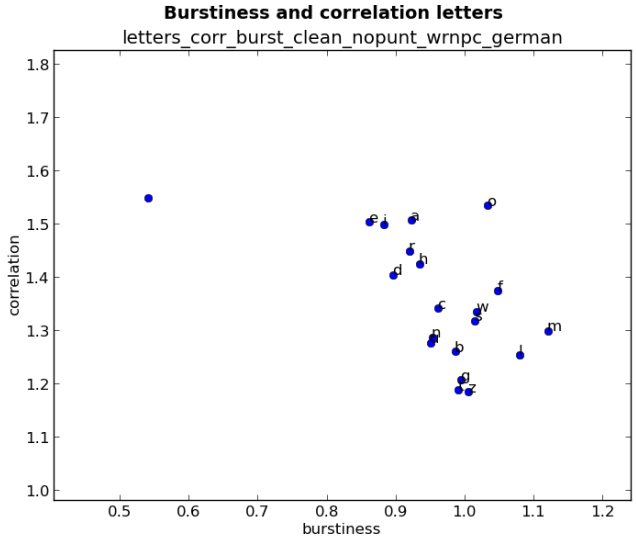



Figure 4.22: Burstiness-Correlation diagram for all symbols studied in *War and Peace* in German.

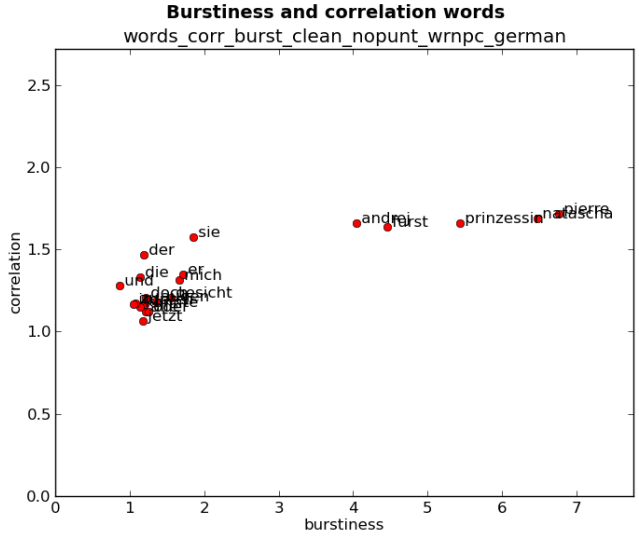



Figure 4.23: Burstiness-Correlation diagram for all words studied in *War and Peace* in German.

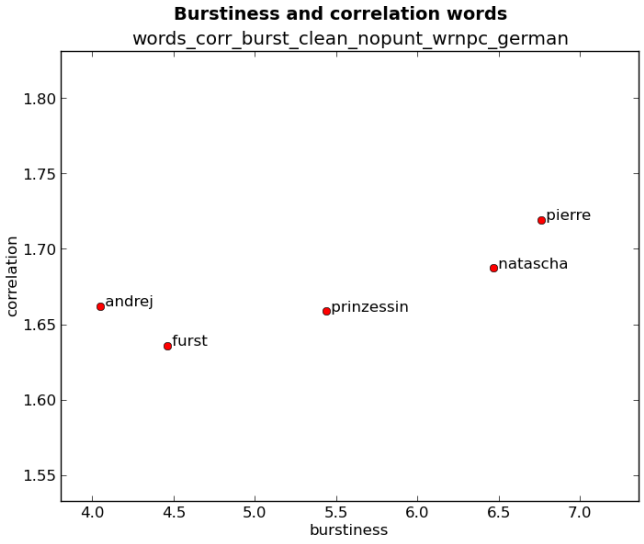



Figure 4.24: Burstiness-Correlation diagram for those words studied in *War and Peace* in German with high values of $B'$ and $\hat{\gamma}$.

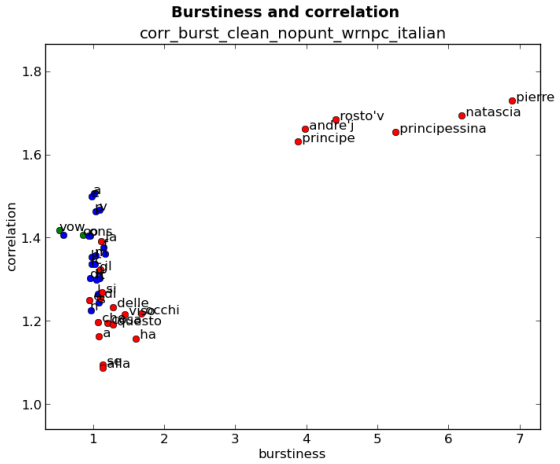



Figure 4.25: Burstiness-Correlation diagram for all 43 binary sequences studied in *War and Peace* in Italian. Green points are vowels (vow) and consonants (cons), blue points are symbols and red points are words.

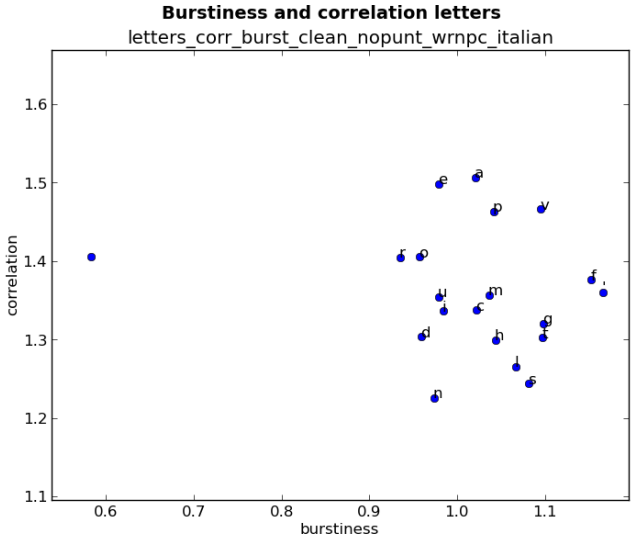



Figure 4.26: Burstiness-Correlation diagram for all symbols studied in *War and Peace* in Italian.

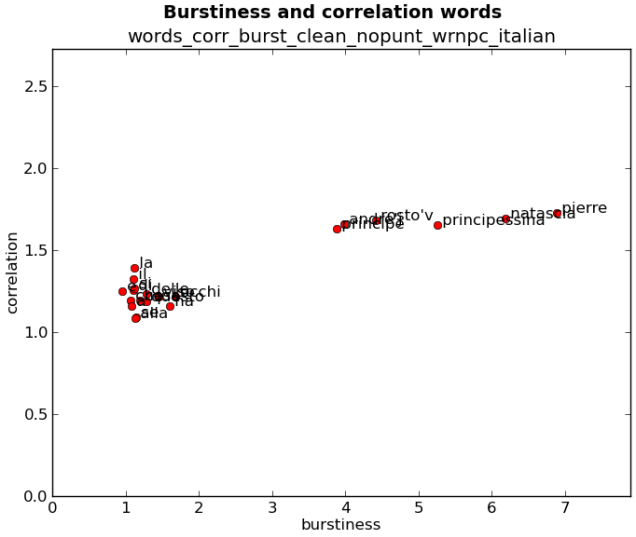



Figure 4.27: Burstiness-Correlation diagram for all words studied in *War and Peace* in Italian.

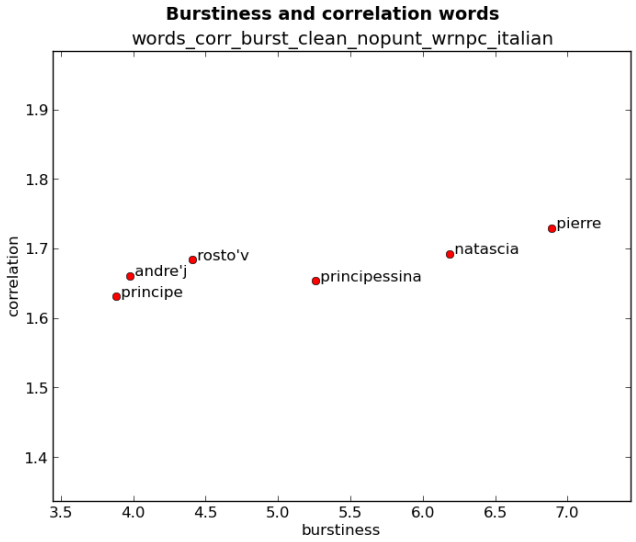



Figure 4.28: Burstiness-Correlation diagram for those words studied in *War and Peace* in Italian with high values of $B'$ and $\hat{\gamma}$.

As we can see in the previous figures and tabs, all letters have $B' \approx 1$, $\hat{\gamma} > 1$ and $\hat{\gamma_2} \approx 1$. This means that correlations is due to $C_\tau(k)$ and not to burstiness. But the most interesting situation takes place in the level of words. The most frequent words show $B' \approx 1$, $\hat{\gamma} > 1$ and $\hat{\gamma_2} \approx 1$ so, as in letter case, correlations is due to $C_\tau(k)$ and not to burstiness. On the contrary, the most frequent nouns with high values of $B'$ show also high values of $\hat{\gamma}$ and of $\hat{\gamma_2}$; the word *" prince "*, studied in the previous section, is an example of this kind of words, for which burstiness strongly influence correlations. In fact, we can easily note that $B'($*" prince "*$) = 3.9026$, $\hat{\gamma}($*" prince "*$) = 1.6324$ and $\hat{\gamma_2}($*" prince "*$) = 1.4101$.

Another important observation we can do is that, contrary to our expectations, the so-called "key-words" reach higher values of $\hat{\gamma}$ than letters ($\hat{\gamma}_{\rm e} < \hat{\gamma}_{\rm prince}$). This fact contradicts the asymptotic behavior studied in the previous chapter: *" prince "* is on top of *"e"* and, from Eq. (3.23), we should have $\hat{\gamma}_{\rm e} \geq \hat{\gamma}_{\rm prince}$. Anyway this seeming contradiction can be easily solved by the estimation of the transition time $t_T$ necessary for the finite time estimate $\hat{\gamma}$ to reach the asymptotic $\gamma$, Eq. (3.34). We can imagine a substitute sequence with the same frequency of *"e"* composed by *" prince "*, with a random addition of ones. Using the fitting values of $g$, $\gamma$ for prince in Eq. (3.34) we obtain $t_T \geq 6 \times 10^5$, larger than the maximum time $t_s$ used to obtain $\hat{\gamma}$. Vice-versa, for a sequence with the same frequency of *" prince "* built as a random sequence on top of *"e"* we obtain $t_T \geq 7 \times 10^8$. These results don't explain the reason why $\hat{\gamma}_{\rm e} < \hat{\gamma}_{\rm prince}$ but we can argue that *" prince "* is a particular meaningful (not random) sequence on top of *"e"* and that *"e"* must necessarily be composed by other sequences with $1 < \gamma < \hat{\gamma}_{\rm prince}$ which dominate for short times. The presence of these sequences also explains the reason why keywords show sharper transitions then letters, as we can easily note in Figures 4.1-4.9.

## 4.3 Combined analysis on languages

Up to now we have analyzed, language per language, long-range correlations and burstiness for a defined set of words. Now we will focus our study on differences and similarity between different languages, using, as above, *War and Peace* in English, French, German and Italian.

Now I will present some of the results obtained from the experiment I did for the study of the comparison of $B'$ and $\hat{\gamma}$ in different languages.

For each language I choose, two keywords (*" prince "* and *" pierre "*), two frequent words (*" and "* and *" in "*), two frequent symbols (" " and *"e"*) and vowels-constants and my main goal is to discover if $B'$ and $\hat{\gamma}$ depend on language or not.

In the following tabs are presented the results of this study for these words-symbols-sequences.

**" prince "**

| word and language | $B'$ | $\hat{\gamma}$ | $\hat{\gamma_2}$ |
|---|---|---|---|
| " prince " English | 3.9026 | 1.6324 | 1.4101 |
| " prince " French | 3.2224 | 1.5819 | 1.4204 |
| " fürst " German | 4.4620 | 1.636 | 1.5006 |
| " principe " Italian | 3.8833 | 1.6309 | 1.4323 |
| Average ($\mu$) | 3.8676 | 1.6203 | 1.4408 |
| Standard deviation ($\sigma$) | 0.43906 | 0.022248 | 0.035380 |
| Coefficient of variation $\left(\frac{\sigma}{\mu}\right)$ | 0.11352 | 0.013731 | 0.024555 |

**" pierre "**

| word and language | $B'$ | $\hat{\gamma}$ | $\hat{\gamma_2}$ |
|---|---|---|---|
| " pierre " English | 6.8589 | 1.7244 | 1.642 |
| " pierre " French | 4.4686 | 1.6915 | 1.5784 |
| " pierre " German | 6.7635 | 1.7192 | 1.626 |
| " pierre " Italian | 6.8904 | 1.7295 | 1.6456 |
| Average ($\mu$) | 6.2454 | 1.7161 | 1.6230 |
| Standard deviation ($\sigma$) | 1.0269 | 0.014690 | 0.026786 |
| Coefficient of variation $\left(\frac{\sigma}{\mu}\right)$ | 0.16442 | 0.0085600 | 0.016504 |

**" and "**

| word and language | $B'$ | $\hat{\gamma}$ | $\hat{\gamma_2}$ |
|---|---|---|---|
| " and " English | 0.88130 | 1.1965 | 0.93914 |
| " et " French | 0.82029 | 1.1889 | 1.018 |
| " und " German | 0.86393 | 1.2799 | 1.0674 |
| " e " Italian | 0.94976 | 1.2496 | 0.97503 |
| Average ($\mu$) | 0.87882 | 1.2287 | 0.99989 |
| Standard deviation ($\sigma$) | 0.046600 | 0.037680 | 0.047943 |
| Coefficient of variation $\left(\frac{\sigma}{\mu}\right)$ | 0.053026 | 0.030666 | 0.047948 |

**" in "**

| word and language | $B'$ | $\hat{\gamma}$ | $\hat{\gamma_2}$ |
|---|---|---|---|
| " in " English | 1.0436 | 1.1471 | 0.95459 |
| " en " French | 1.0381 | 1.1724 | 1.0272 |
| " in " German | 1.0558 | 1.1643 | 0.95713 |
| " in " Italian | 1.0674 | 1.0711 | 0.956 |
| Average ($\mu$) | 1.0512 | 1.138725 | 0.97373 |
| Standard deviation ($\sigma$) | 0.011314 | 0.040098 | 0.030884 |
| Coefficient of variation $\left(\frac{\sigma}{\mu}\right)$ | 0.010763 | 0.035213 | 0.031717 |

**" "**

| word and language | $B'$ | $\hat{\gamma}$ | $\hat{\gamma_2}$ |
|---|---|---|---|
| " " English | 0.52981 | 1.5052 | 0.9785 |
| " " French | 0.61263 | 1.3967 | 0.97946 |
| " " German | 0.54066 | 1.5489 | 0.99867 |
| " " Italian | 0.58245 | 1.4059 | 1.0064 |
| Average ($\mu$) | 0.56639 | 1.4642 | 0.99076 |
| Standard deviation ($\sigma$) | 0.033150 | 0.064827 | 0.012095 |
| Coefficient of variation $\left(\frac{\sigma}{\mu}\right)$ | 0.058530 | 0.044276 | 0.012208 |

**"e"**

| word and language | $B'$ | $\hat{\gamma}$ | $\hat{\gamma_2}$ |
|---|---|---|---|
| "e" English | 0.92834 | 1.3738 | 1.0211 |
| "e" French | 0.86628 | 1.3551 | 1.0257 |
| "e" German | 0.86223 | 1.5034 | 1.0129 |
| "e" Italian | 0.97890 | 1.4982 | 1.0564 |
| Average ($\mu$) | 0.90894 | 1.4326 | 1.0290 |
| Standard deviation ($\sigma$) | 0.048148 | 0.068519 | 0.016457 |
| Coefficient of variation $\left(\frac{\sigma}{\mu}\right)$ | 0.052972 | 0.047828 | 0.015992 |

**vowels**

| word and language | $B'$ | $\hat{\gamma}$ | $\hat{\gamma_2}$ |
|---|---|---|---|
| vowels English | 0.61797 | 1.5271 | 1.007 |
| vowels French | 0.67026 | 1.321 | 0.97224 |
| vowels German | 0.62397 | 1.3008 | 1.0509 |
| vowels Italian | 0.53241 | 1.4171 | 0.94296 |
| Average ($\mu$) | 0.61115 | 1.3915 | 0.99327 |
| Standard deviation ($\sigma$) | 0.049761 | 0.089777 | 0.040259 |
| Coefficient of variation $\left(\frac{\sigma}{\mu}\right)$ | 0.081421 | 0.064518 | 0.040531 |

**consonants**

| word and language | $B'$ | $\hat{\gamma}$ | $\hat{\gamma}_2$ |
|---|---|---|---|
| consonants English | 0.83332 | 1.445 | 1.0461 |
| consonants French | 0.79767 | 1.3792 | 1.0209 |
| consonants German | 0.89555 | 1.3805 | 1.033 |
| consonants Italian | 0.85645 | 1.4055 | 1.0171 |
| Average ($\mu$) | 0.84575 | 1.4026 | 1.0293 |
| Standard deviation ($\sigma$) | 0.035568 | 0.026656 | 0.011350 |
| Coefficient of variation $\left(\frac{\sigma}{\mu}\right)$ | 0.042055 | 0.019005 | 0.11028 |

These results can be better intrepreted thanks to the following figures, where there are three different views of a $B' - \hat{\gamma}$ graph for all these words-symbols-sequences together and singular graphs for each sequence. In the latter case in the graphs there are also rectangles with these vertices:

$$(\mu_{B'} - \sigma_{B'}, \mu_\gamma - \sigma_\gamma)\,,\ (\mu_{B'} - \sigma_{B'}, \mu_\gamma + \sigma_\gamma)\,,$$

$$(\mu_{B'} + \sigma_{B'}, \mu_\gamma + \sigma_\gamma) \text{ and } (\mu_{B'} + \sigma_{B'}, \mu_\gamma - \sigma_\gamma)$$

where $\mu$ is the average and $\sigma$ is the standard deviation.

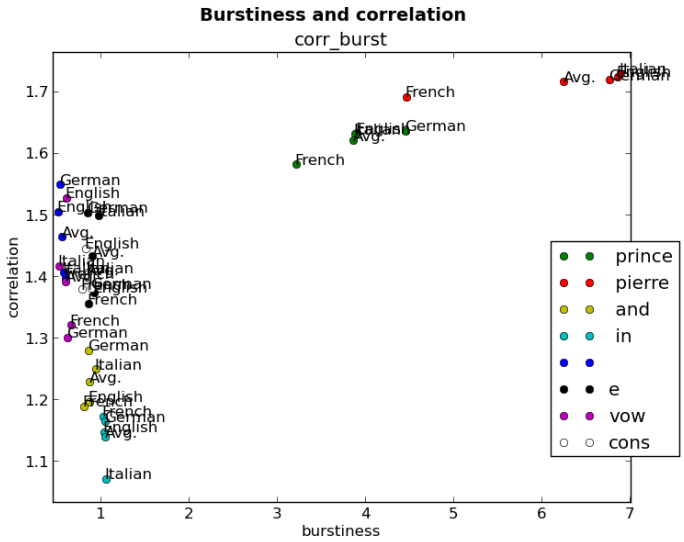



Figure 4.29: Burstiness-Correlation diagram for all those sequences studied in *War and Peace* in all languages.

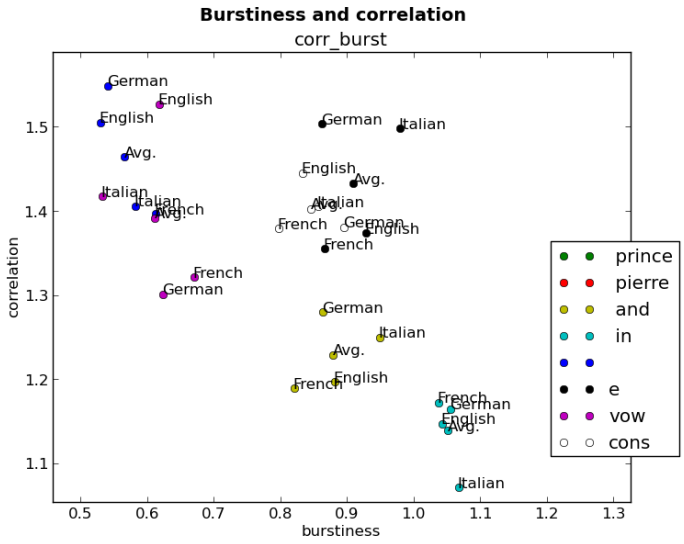



Figure 4.30: Burstiness-Correlation diagram for those sequences studied in *War and Peace* in all languages with low values of $B'$.

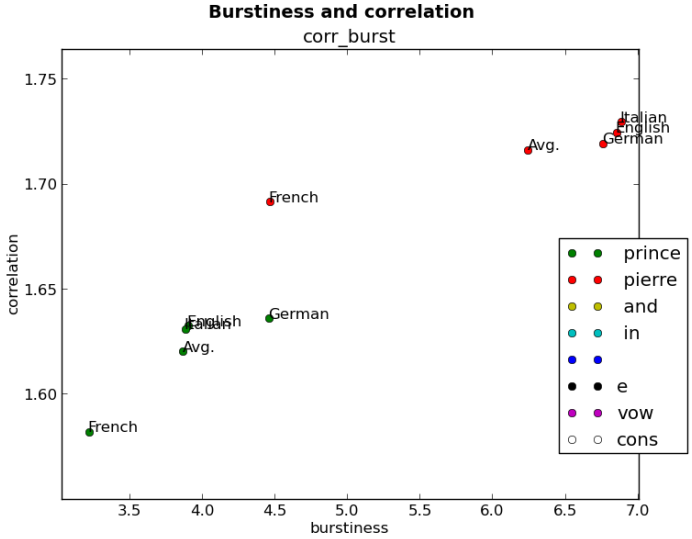



Figure 4.31: Burstiness-Correlation diagram for all those sequences studied in *War and Peace* in all languages with high values of $B'$.

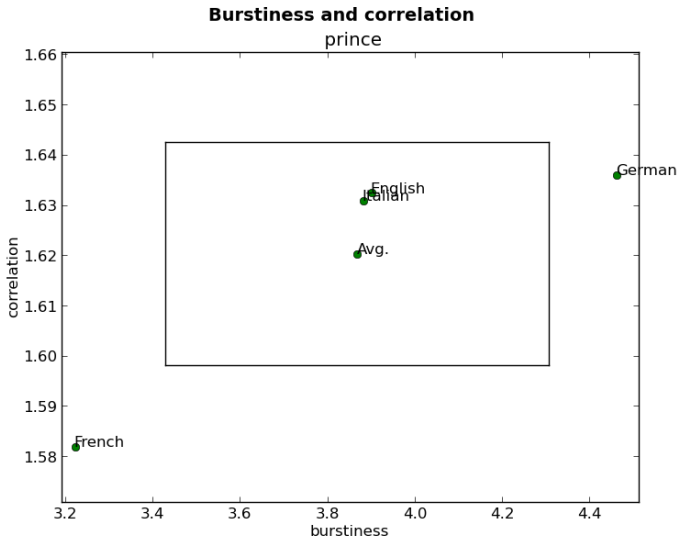



Figure 4.32: Burstiness-Correlation diagram for the word *" prince "*, studied in *War and Peace* in all languages.

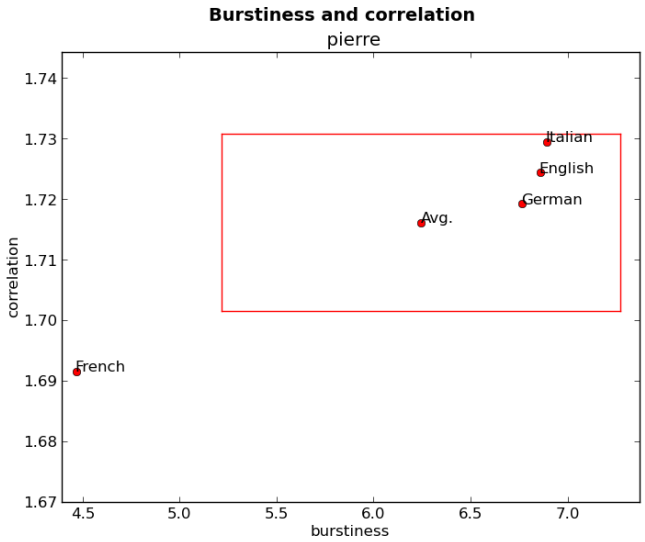



Figure 4.33: Burstiness-Correlation diagram for the word *" pierre "*, studied in *War and Peace* in all languages.

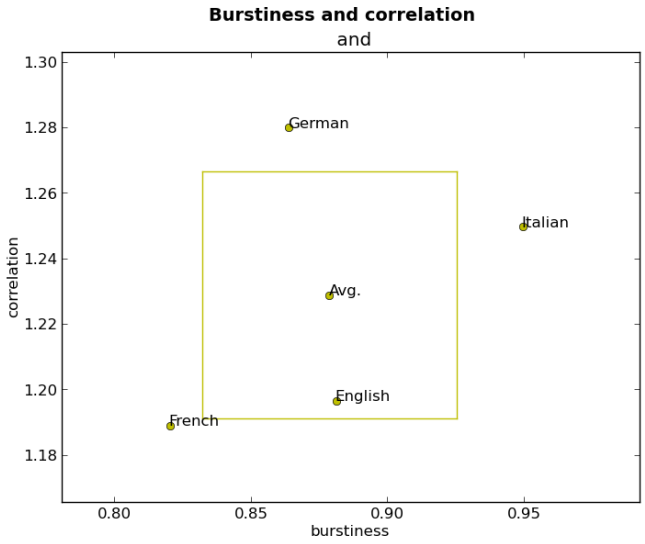
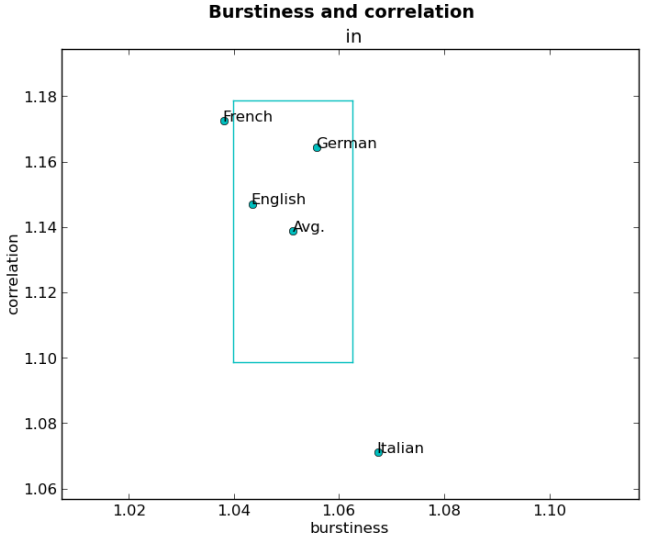



Figure 4.34: Burstiness-Correlation diagram for the word *" and "*, studied in *War and Peace* in all languages.




prince

Figure 4.35: Burstiness-Correlation diagram for the word ” *in* ”, studied in *War and Peace* in all languages.

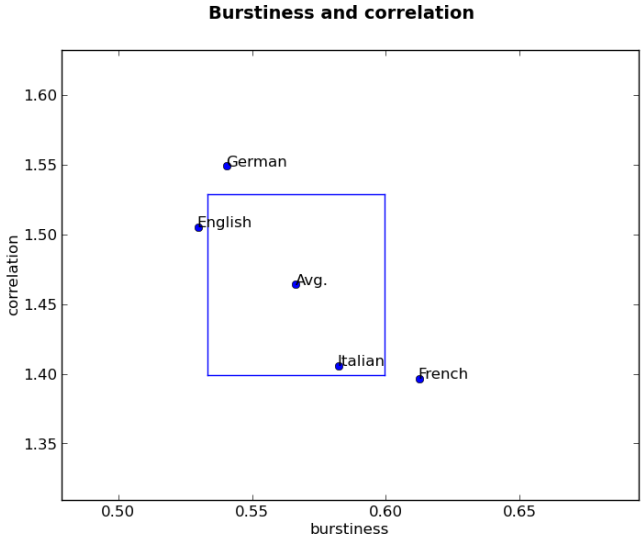



Figure 4.36: Burstiness-Correlation diagram for the space ” ”, studied in *War and Peace* in all languages.

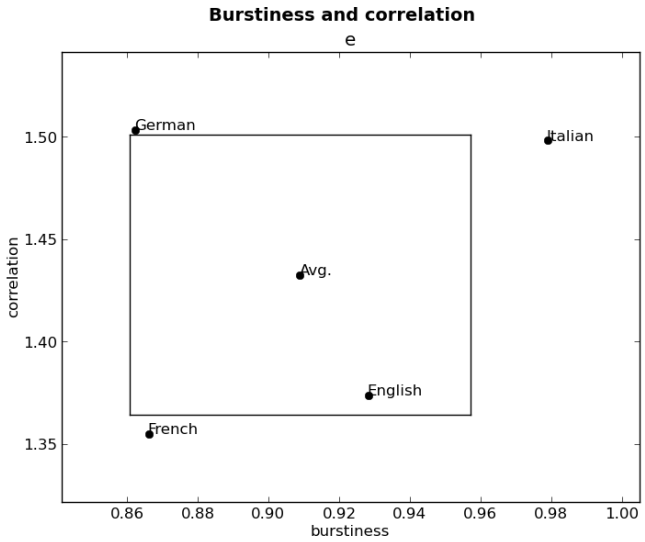



Figure 4.37: Burstiness-Correlation diagram for the symbol *"e"*, studied in *War and Peace* in all languages.

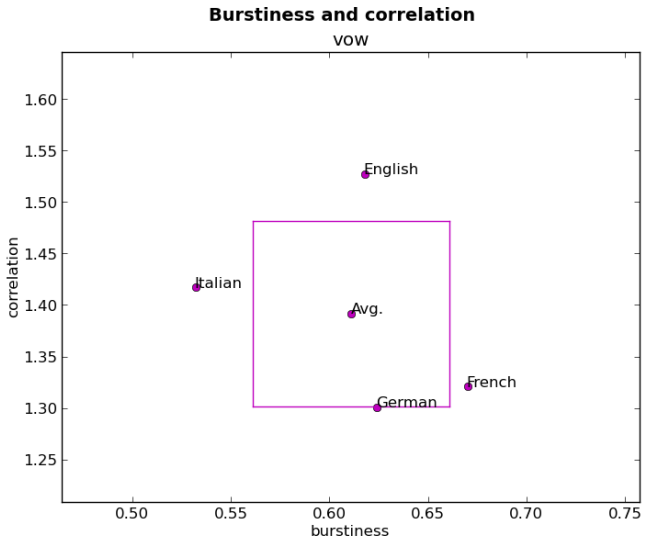



Figure 4.38: Burstiness-Correlation diagram for the vowels, studied in *War and Peace* in all languages.

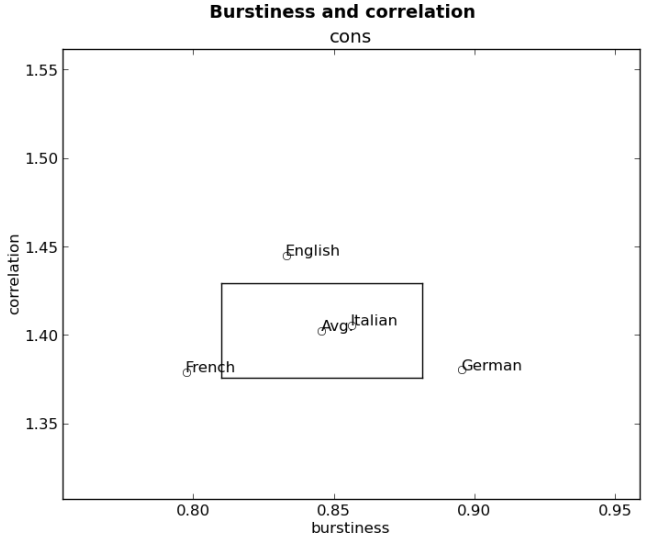



Figure 4.39: Burstiness-Correlation diagram for the consonants, studied in *War and Peace* in all languages.

As we can see from the previous figures and tabs, burstiness coefficient, especially for key-words, isn't conserved in different languages, while, on the contrary, long-range correlations exponent seems to be preserved precisely for key-words. In fact $\sigma_\gamma$ for " *prince* " and " *pierre* " is smaller than $\sigma_\gamma$ for all the others sequences. So we may argue that, while letters, vowels-consonants and "not key-words", before reaching their asymptotic behavior, are more influenced by the used language, key-words, reaching their asymptotic behavior earlier, behave in the same way, without a strong dependence on the language chosen. An hypothesis for this result may be that, if we work with enough "short" sequences, the influence on long-range correlations of the burstiness doesn't depend on language, while the influence on long-range correlations of $C_\tau(k)$ depends on language and the obtained values of $\hat{\gamma}_2$ seem to confirm this idea. Obviously, this hypothesis has to be tested with other experiments. In order to confirm or reject this hypothesis we may repeat this experiment using other books, like *The Bible*.

Before closing the chapter there is an apparently strange result observable that I'd like to explain: the value of $B'$ for the word *" pierre "* in French. In fact

$$B'_{\text{" pierre "}}(\text{English}) = 6.8589, \tag{4.2}$$

$$B'_{\text{" pierre "}}(\text{French}) = 4.4686, \tag{4.3}$$

$$B'_{\text{" pierre "}}(\text{German}) = 6.7635, \tag{4.4}$$

$$B'_{\text{" pierre "}}(\text{Italian}) = 6.8904. \tag{4.5}$$

Hence

$$\mu(B'_{\text{" pierre "}}) = 6.2454, \tag{4.6}$$

$$\sigma(B'_{\text{" pierre "}}) = 1.0269. \tag{4.7}$$

It seems that the sequence *" pierre "* in French follows a behavior completely different from the other languages; but this strange value can be easily solved searching this word in the text and observing that the word *" pierre "* means both *Pierre*, a character of the book, and *pierre*, that means stone. For example the following paragraph presents this particular fact.

*"Marche! marche! Trois roubles de pourboire!" s'écria Rostow, qui, à quelques pas de chez lui, croyait ne jamais arriver. Le traîneau prit sur la droite et s'arrêta devant le perron. Rostow reconnut la corniche ébréchée, la borne du trottoir, et s'élança hors du traîneau avant qu'il se fût arrûté. Il franchit les marches d'un bond. L'extérieur de la maison était aussi froid, aussi calme que par le passé. Que faisait à ces murs de* ***pierre*** *l'arrivée ou le départ? Personne dans le vestibule! "Mon Dieu! serait-il arrivé quelque chose?" se dit Rostow avec un serrement de coeur; il s'arrêta une minute, puis reprit sa course dans l'escalier aux marches usées, qu'il connaissait si bien. "Et voilà le même bouton de porte déjeté, dont la malpropreté agaçait toujours la comtesse, et voilà l'antichambre!" Elle n'était éclairée dans ce moment que par une chandelle.*

In English the translation is:

*"Then they've not gone to bed yet? What do you think? Mind now, don't forget to put out my new coat," added Rostov, fingering his new mustache. "Now then, get on," he shouted to the driver. "Do wake up, Vaska!" he went on, turning to Denisov, whose head was again nodding. "Come, get on! You shall have three rubles for vodka–get on!" Rostov shouted, when the sleigh was only three houses from his door. It seemed to him the horses were not moving at all. At last the sleigh bore to the right, drew up at an entrance, and Rostov saw overhead the old familiar cornice with a bit of plaster broken off, the porch, and the post by the side of the pavement. He sprang out before the sleigh stopped, and ran into the hall. The house stood cold and silent, as if quite regardless of who had come to it. There was no one in the hall. "Oh God! Is everyone all right?" he thought, stopping for a moment with a sinking heart, and then immediately starting to run along the hall and up the warped steps of the familiar staircase. The well-known old door handle, which always angered the countess when it was not properly cleaned, turned as loosely as ever. A solitary tallow candle burned in the anteroom.*

# Chapter 5

# Comparison with another approach for text analysis

In my work of thesis I focused on a particular approach for text analysis. Two colleagues of mine, Filippo Bonora and Giulia Tini, who prepared their master theses with Mirko Degli Esposti and Giampaolo Cristadoro at the same time with me, used a completely different approach for text analysis, that is considering a text as a network. In the first section there will be a brief introduction to their approach (for a more detailed explanation see Filippo's and Giulia's theses), while in the second one there will be a brief comparison between results obtained using these two different approaches, focusing our analysis on key-words. This second section has been written together with Filippo and Giulia.

## 5.1 Texts as networks

This approach is based on the idea that a graph is a very interesting way to describe interactions between words in a text. A useful opportunity to build it from a text is to associate a vertex to each sign of the text (words and punctuation) and to put a link between two vertices if they are adjacent in the text. Hence this approach uses networks analyses to investigate and

discover features of texts.

First of all, here there are some basic definitions of network theory.

**Definition 5.1.1.** A weighted directed graph $G$ is defined by:

- a set $N(G)$ of $N$ vertices, or nodes, identified by an integer value, $i = 1, 2, ..., N$;
- a set $E(G)$ of $M$ edges, or links, identified by a pair $(i, j)$ that represents a connection starting in vertex $i$ and going to vertex $j$.
- a mapping $\omega : E(G) \to \mathbb{R}$ that associate to the edge $(i, j)$ the value $\omega(i, j) = \omega_{i,j}$ called weight.

**Definition 5.1.2.** A weighted directed graph $G$ can be represented using its weight matrix $W = (\omega_{i,j})$, an $N \times N$ matrix whose elements represent the number of directed links connecting vertex $i$ with vertex $j$. In this work we will assume that no pair of edges $(i_1, j_1)$ and $(i_2, j_2)$ with $i_1 = i_2$ and $j_1 = j_2$ exist.

**Definition 5.1.3.** The $N \times N$ matrix $A = (a_{i,j})$ is the adjacency matrix of the graph $G$ if

$$\forall i, j, a_{i,j} = \begin{cases} 1, & \text{if } \omega_{i,j} \neq 0 \\ 0, & \text{if } \omega_{i,j} = 0 \end{cases}$$

**Definition 5.1.4.** The neighborhood of a vertex $i$, $\nu_i$, is the set of vertices adjacent to $i$.

$$\nu_i = \{j \in N(G) : (i, j) \vee (j, i) \in E(G)\}$$

**Definition 5.1.5.** Eventually two non adjacent vertices $i$ and $j$ can be connected using a walk, that is a sequence of $m$ edges $(i, k_1)$, $(k_1, k_2)$, ..., $(k_{m-1}, j)$. If all the edges and all the vertices composing a walk are distinct, the walk is called path.

**Definition 5.1.6.** A shortest path between two nodes is defined as the path whose sum of edge weights is minimum.

Given these definitions, we can compute the importance of a vertex or an edge considering the number of paths in which it is involved and, assuming that a vertex is reached using the shortest path, this can be measured by the betweenness centrality.

**Definition 5.1.7.** The betweenness centrality of a vertex or an edge $u$ is defined as

$$B_u = \sum_{i,j} \frac{\sigma(i,u,j)}{i,j},$$

where $\sigma(i,u,j)$ is the number of shortest paths between vertices $i$ and $j$ that pass through $u$ while $\sigma(i,j)$ is the total number of shortest paths between $i$ and $j$.

Hence, once built the graph associated to the text, we can consider as keywords these words with the highest values of betweenness centrality. But, as explained in the theses of Filippo Bonora and of Giulia Tini, in order to obtain good results, it's fundamental cleaning the text from stopwords (articles, preposition, adverbs, ...). In their analysis they built graphs from texts with Python and made statistical analysis on graphs with Gephi, an open-source software for visualizing and analyzing large networks graphs.

## 5.2 Comparison of results

For the comparison of results we used the following books: in English *Moby Dick* by Herman Melville, *A naturalist's voyage round the world* by Charles Darwin, *Alice's adventures in wonderland* by Lewis Carroll, *Life on Mississippi* by Mark Twain and *The jungle* by Sinclair Upton; in Italian *I Malavoglia* by Giovanni Verga, *I pirati della Malesia* by Emilio Salgari, *Le avventure di Pinocchio-Storia di un burattino* by Carlo Collodi, *Canne al vento* by Grazia Deledda, *Il fu Mattia Pascal* by Luigi Pirandello and *La coscienza di Zeno* by Italo Svevo.

Comparing the words with highest betweenness centrality, found by Filippo and Giulia, with the words studied with the procedure explained in Chapter 4, although there are some differences, we can see also many similarities, especially for the longest texts used in this particular analysis.

Before going on with this analysis, it's important to underline that, while in my method words are chosen with a particular procedure, Filippo's and Giulia's approach is much more general, because betweenness centrality is calculated for all words present in the analyzed text, cleaned from stopwords.

In the following tabs there are results obtained using my approach.

***Moby Dick*, N=1151326**

| $\alpha$ | $f$ | $B'$ | $\hat{\gamma}$ | $\sigma_{\hat{\gamma}} \times 10^3$ |
|---|---|---|---|---|
| " whale " | 1150 | 2.1503 | 1.555 | 0.67814 |
| " not " | 1142 | 1.0936 | 1.0577 | 0.18287 |
| " man " | 525 | 1.2805 | 1.1944 | 0.25952 |
| " into " | 517 | 1.1829 | 1.1771 | 0.42026 |
| " ahab " | 510 | 3.2174 | 1.5502 | 0.64730 |
| " ship " | 509 | 1.5666 | 1.309 | 0.37457 |
| " up " | 508 | 1.1642 | 1.1813 | 0.34409 |
| " more " | 503 | 1.1474 | 1.0542 | 0.090317 |
| " sea " | 437 | 1.3010 | 1.1959 | 0.32605 |
| " would " | 426 | 1.1825 | 1.0774 | 0.34327 |
| " head " | 337 | 1.3334 | 1.3466 | 0.41653 |
| " time " | 332 | 1.0881 | 1.0505 | 0.18922 |
| " boat " | 331 | 2.4121 | 1.4585 | 0.36717 |
| " her " | 330 | 1.9006 | 1.2136 | 0.34803 |
| " the " | 14168 | 1.0453 | 1.384 | 0.27379 |
| " of " | 6469 | 1.0779 | 1.4589 | 0.62232 |
| " and " | 6325 | 0.96824 | 1.2176 | 0.33535 |
| " a " | 4630 | 1.1417 | 1.3306 | 0.41654 |
| " to " | 4539 | 1.0233 | 1.1348 | 0.11508 |
| " in " | 4076 | 1.0230 | 1.2816 | 0.57843 |
| " that " | 3037 | 1.0974 | 1.1397 | 0.15402 |

***A naturalist's voyage round the world*, N=1153329**

| $\alpha$ | $f$ | $B'$ | $\hat{\gamma}$ | $\sigma_{\hat{\gamma}} \times 10^3$ |
|---|---|---|---|---|
| " water " | 426 | 1.5209 | 1.2276 | 0.14963 |
| " little " | 412 | 1.1178 | 1.0691 | 0.20576 |
| " sea " | 348 | 1.5347 | 1.2571 | 0.23099 |
| " up " | 340 | 1.1912 | 1.1203 | 0.40453 |
| " country " | 337 | 1.5194 | 1.2534 | 0.37560 |
| " being " | 328 | 1.0600 | 1.0659 | 0.37220 |
| " day " | 327 | 1.4439 | 1.1378 | 0.26183 |
| " land " | 318 | 1.3876 | 1.2379 | 0.45840 |
| " must " | 317 | 1.2906 | 1.1112 | 0.37057 |
| " them " | 315 | 1.1098 | 1.1011 | 0.19440 |
| " feet " | 312 | 1.3912 | 1.2036 | 0.12373 |
| " may " | 311 | 1.1184 | 1.0731 | 0.18825 |
| " species " | 303 | 2.4601 | 1.5192 | 0.25985 |
| " if " | 302 | 1.1387 | 1.0955 | 0.21750 |
| " the " | 16878 | 0.93709 | 1.1878 | 0.067457 |
| " of " | 9411 | 0.97763 | 1.2284 | 0.15445 |
| " and " | 5762 | 0.90143 | 1.0791 | 0.068272 |
| " a " | 5333 | 1.1026 | 1.1689 | 0.17525 |
| " in " | 4287 | 1.0255 | 1.1297 | 0.32472 |
| " to " | 4080 | 1.0543 | 1.1863 | 0.29505 |
| " is " | 2414 | 1.3143 | 1.2054 | 0.16060 |

***Alice's adventures in wonderland*, N=135006**

| $\alpha$ | $f$ | $B'$ | $\hat{\gamma}$ | $\sigma_{\hat{\gamma}} \times 10^3$ |
|---|---|---|---|---|
| " alice " | 397 | 0.90337 | 1.0926 | 1.7490 |
| " in " | 368 | 0.97125 | 1.0423 | 0.10162 |
| " queen " | 75 | 2.1158 | 1.4876 | 1.6266 |
| " thought " | 74 | 0.91569 | 0.98403 | 0.60826 |
| " time " | 71 | 0.97476 | 1.1338 | 1.0262 |
| " how " | 68 | 1.1639 | 1.0911 | 0.38924 |
| " king " | 63 | 2.7741 | 1.5343 | 1.7592 |
| " your " | 62 | 1.3422 | 1.1088 | 1.0116 |
| " turtle " | 59 | 3.6274 | 1.6217 | 1.9307 |
| " my " | 58 | 1.0647 | 1.1319 | 0.59194 |
| " way " | 56 | 1.0984 | 1.2266 | 0.19224 |
| " mock " | 56 | 3.5250 | 1.6079 | 2.0555 |
| " hatter " | 56 | 5.0412 | 1.6351 | 1.2570 |
| " quite " | 55 | 1.2917 | 1.1747 | 0.55709 |
| " the " | 1641 | 0.98188 | 1.3563 | 0.78243 |
| " and " | 871 | 0.96193 | 1.1100 | 0.97739 |
| " to " | 729 | 1.01958 | 1.0958 | 0.74467 |
| " a " | 632 | 1.0767 | 1.0732 | 0.24094 |
| " it " | 595 | 1.1694 | 1.24464 | 0.24464 |
| " she " | 552 | 1.5668 | 1.3817 | 0.90159 |
| " i " | 544 | 1.5967 | 1.3174 | 0.12501 |

***Life on Mississippi*, N=767745**

| $\alpha$ | $f$ | $B'$ | $\hat{\gamma}$ | $\sigma_{\hat{\gamma}} \times 10^3$ |
|---|---|---|---|---|
| " river " | 489 | 2.2193 | 1.4259 | 0.63319 |
| " which " | 482 | 1.2217 | 1.1564 | 0.27336 |
| " time " | 355 | 1.2064 | 1.0932 | 0.54676 |
| " down " | 341 | 1.3091 | 1.2082 | 0.39035 |
| " man " | 278 | 1.6388 | 1.2553 | 0.42158 |
| " its " | 276 | 1.5620 | 1.303 | 0.38111 |
| " water " | 246 | 1.9250 | 1.3436 | 0.26651 |
| " got " | 234 | 1.5849 | 1.2294 | 0.59812 |
| " boat " | 234 | 2.0364 | 1.3421 | 0.67421 |
| " these " | 231 | 1.5343 | 1.1751 | 0.25742 |
| " day " | 224 | 1.1676 | 1.0052 | 0.21229 |
| " can " | 219 | 1.4214 | 1.1694 | 0.12941 |
| " way " | 217 | 1.0263 | 1.0284 | 0.20210 |
| " did " | 216 | 1.5945 | 1.1737 | 0.54900 |
| " the " | 9043 | 1.0510 | 1.355 | 0.19181 |
| " and " | 5879 | 0.97594 | 1.3217 | 0.63667 |
| " of " | 4363 | 1.0372 | 1.3084 | 0.58920 |
| " a " | 4049 | 1.0971 | 1.1951 | 0.40830 |
| " to " | 3531 | 1.0989 | 1.2252 | 0.35808 |
| " in " | 2535 | 1.0340 | 1.1215 | 0.24779 |
| " it " | 2369 | 1.3739 | 1.2647 | 0.16832 |

***The jungle*, N=783190**

| $\alpha$ | $f$ | $B'$ | $\hat{\gamma}$ | $\sigma_{\hat{\gamma}} \times 10^3$ |
|---|---|---|---|---|
| ” jurgis ” | 1117 | 2.0933 | 1.4772 | 0.80927 |
| ” were ” | 995 | 1.5754 | 1.3132 | 0.40206 |
| ” man ” | 481 | 1.3175 | 1.2682 | 0.45402 |
| ” an ” | 436 | 1.2217 | 1.2281 | 0.63330 |
| ” time ” | 358 | 1.1973 | 1.1396 | 0.47247 |
| ” if ” | 346 | 1.1691 | 1.0852 | 0.31373 |
| ” men ” | 340 | 1.7881 | 1.3121 | 0.44895 |
| ” now ” | 325 | 1.0770 | 1.0938 | 0.19913 |
| ” day ” | 288 | 1.3975 | 1.1381 | 0.42269 |
| ” get ” | 279 | 1.3328 | 1.1444 | 0.38899 |
| ” place ” | 263 | 1.2274 | 1.1505 | 0.33420 |
| ” like ” | 261 | 1.0561 | 1.0432 | 0.18738 |
| ” home ” | 229 | 1.7595 | 1.2163 | 0.35910 |
| ” ona ” | 225 | 3.7697 | 1.4446 | 0.60946 |
| ” the ” | 8925 | 1.0286 | 1.3068 | 0.23050 |
| ” and ” | 7260 | 0.96796 | 1.2425 | 0.20601 |
| ” of ” | 4364 | 1.1190 | 1.4178 | 0.85064 |
| ” to ” | 4187 | 1.0820 | 1.1979 | 0.60636 |
| ” a ” | 4160 | 1.1554 | 1.2047 | 0.25191 |
| ” he ” | 3310 | 2.1662 | 1.5723 | 0.30364 |
| ” was ” | 3055 | 1.8421 | 1.4281 | 0.87478 |

***I Malavoglia*, N=488279**

| $\alpha$ | $f$ | $B'$ | $\hat{\gamma}$ | $\sigma_{\hat{\gamma}} \times 10^3$ |
|---|---|---|---|---|
| " ntoni " | 557 | 1.5914 | 1.3578 | 1.1354 |
| " aveva " | 480 | 1.2099 | 1.0874 | 0.40255 |
| " don " | 431 | 2.4042 | 1.4901 | 0.87779 |
| " piu' " | 425 | 1.1696 | 1.1467 | 0.43453 |
| " casa " | 361 | 1.2793 | 1.1593 | 0.30674 |
| " ma " | 355 | 0.99082 | 1.0931 | 0.69757 |
| " padron " | 333 | 1.5638 | 1.2493 | 0.44537 |
| " una " | 332 | 0.95941 | 0.96133 | 0.42429 |
| " zio " | 194 | 1.7520 | 1.3398 | 0.88481 |
| " compare " | 194 | 1.6203 | 1.1773 | 0.33540 |
| " o " | 194 | 1.1081 | 1.0638 | 0.14430 |
| " quale " | 191 | 1.1073 | 1.0187 | 0.25939 |
| " malavoglia " | 191 | 1.2946 | 1.1419 | 0.48162 |
| " cosa " | 190 | 1.1135 | 1.0789 | 0.33912 |
| " e " | 3487 | 0.85760 | 1.1211 | 0.42961 |
| " che " | 2554 | 0.84766 | 1.0051 | 0.27934 |
| " la " | 2310 | 1.0752 | 1.193 | 0.63733 |
| " a " | 1996 | 0.97622 | 1.1233 | 0.68903 |
| " di " | 1977 | 1.0824 | 1.0966 | 0.45521 |
| " il " | 1873 | 1.0287 | 1.1283 | 0.27638 |
| " non " | 1719 | 1.0158 | 1.1636 | 0.30869 |

***I pirati della Malesia*, N=349295**

| $\alpha$ | $f$ | $B'$ | $\hat{\gamma}$ | $\sigma_{\hat{\gamma}} \times 10^3$ |
|---|---|---|---|---|
| ” yanez ” | 388 | 2.2748 | 1.3814 | 1.0383 |
| ” da ” | 380 | 1.1378 | 0.99848 | 0.29047 |
| ” sandokan ” | 364 | 2.5877 | 1.5032 | 1.7599 |
| ” disse ” | 350 | 1.3937 | 1.2664 | 1.0460 |
| ” rajah ” | 246 | 2.1324 | 1.4204 | 1.3580 |
| ” tigre ” | 231 | 1.9718 | 1.2757 | 0.50372 |
| ” al ” | 228 | 1.0200 | 1.0289 | 0.18203 |
| ” piu’ ” | 212 | 1.2401 | 1.218 | 0.33829 |
| ” kammamuri ” | 205 | 1.7348 | 1.2721 | 0.73514 |
| ” dei ” | 196 | 1.1197 | 1.0452 | 0.65963 |
| ” malesia ” | 170 | 1.7603 | 1.2092 | 0.79521 |
| ” sono ” | 165 | 1.2613 | 1.0938 | 0.18944 |
| ” pirati ” | 148 | 1.7240 | 1.231 | 0.72050 |
| ” aveva ” | 146 | 1.2428 | 1.1948 | 0.40056 |
| ” di ” | 1767 | 1.0877 | 1.1347 | 0.18108 |
| ” e ” | 1600 | 0.95541 | 1.1746 | 0.73010 |
| ” il ” | 1327 | 1.0633 | 1.1935 | 0.69366 |
| ” che ” | 1263 | 0.99470 | 1.1195 | 0.39583 |
| ” la ” | 1167 | 1.1152 | 1.1666 | 0.36438 |
| ” un ” | 1006 | 1.1359 | 1.1747 | 0.52985 |
| ” a ” | 864 | 1.0902 | 1.094 | 0.16491 |

***Le avventure di Pinocchio-Storia di un burattino*, N=226749**

| $\alpha$ | $f$ | $B'$ | $\hat{\gamma}$ | $\sigma_{\hat{\gamma}} \times 10^3$ |
|---|---|---|---|---|
| " pinocchio " | 416 | 0.90728 | 1.0646 | 0.97828 |
| " si " | 393 | 1.1271 | 1.0528 | 1.0272 |
| " se " | 189 | 0.99442 | 1.0414 | 0.28891 |
| " casa " | 93 | 1.5900 | 1.2095 | 0.94809 |
| " nel " | 93 | 1.1993 | 1.0346 | 0.23003 |
| " fata " | 80 | 1.9789 | 1.3934 | 2.1290 |
| " loro " | 78 | 1.4340 | 1.1269 | 0.78794 |
| " babbo " | 74 | 2.4542 | 1.1293 | 0.65502 |
| " altro " | 74 | 1.1774 | 1.0086 | 0.27525 |
| " cosa " | 73 | 1.1088 | 1.062 | 0.50570 |
| " geppetto " | 72 | 3.9571 | 1.5674 | 2.2448 |
| " ragazzi " | 69 | 1.6898 | 1.1785 | 1.0180 |
| " fu " | 67 | 0.98177 | 1.0365 | 0.28429 |
| " e " | 1763 | 0.89783 | 1.1505 | 0.50405 |
| " di " | 1339 | 1.0314 | 0.99713 | 0.56542 |
| " che " | 1019 | 1.0030 | 1.0491 | 0.20004 |
| " a " | 936 | 0.98894 | 1.0347 | 0.32038 |
| " il " | 925 | 1.0172 | 1.0784 | 0.39999 |
| " un " | 762 | 1.0332 | 0.98896 | 0.32700 |
| " la " | 711 | 1.1404 | 1.1816 | 0.65105 |

***Canne al vento*, N=338621**

| $\alpha$ | $f$ | $B'$ | $\hat{\gamma}$ | $\sigma_{\hat{\gamma}} \times 10^3$ |
|---|---|---|---|---|
| ” efix ” | 502 | 1.7046 | 1.244 | 1.8360 |
| ” l ” | 478 | 1.0633 | 1.0625 | 0.20155 |
| ” donna ” | 281 | 2.4683 | 1.3287 | 0.96382 |
| ” da ” | 267 | 1.0164 | 1.0545 | 0.39254 |
| ” noemi ” | 264 | 2.3610 | 1.4505 | 1.6167 |
| ” della ” | 262 | 1.1563 | 1.0948 | 0.13872 |
| ” don ” | 200 | 1.8383 | 1.2731 | 0.86196 |
| ” giacinto ” | 183 | 2.4431 | 1.3358 | 1.0827 |
| ” occhi ” | 182 | 1.0094 | 1.0178 | 0.15531 |
| ” aveva ” | 177 | 1.0854 | 1.0429 | 0.96563 |
| ” suo ” | 171 | 1.0497 | 1.0492 | 0.20738 |
| ” predu ” | 166 | 2.3467 | 1.407 | 1.1090 |
| ” ed ” | 164 | 1.3342 | 1.0121 | 0.36151 |
| ” e ” | 2244 | 0.91771 | 1.0199 | 0.37359 |
| ” di ” | 1729 | 1.1009 | 1.1428 | 0.26812 |
| ” la ” | 1282 | 1.0560 | 1.0908 | 0.18029 |
| ” il ” | 1249 | 1.0248 | 1.0837 | 0.49536 |
| ” che ” | 1072 | 1.0019 | 1.0942 | 0.41376 |
| ” e’ ” | 934 | 1.8201 | 1.2749 | 0.39913 |
| ” a ” | 917 | 0.98852 | 1.0005 | 0.17806 |

***Il fu Mattia Pascal*, N=433543**

| $\alpha$ | $f$ | $B'$ | $\hat{\gamma}$ | $\sigma_{\hat{\gamma}} \times 10^3$ |
|---|---|---|---|---|
| ” adriana ” | 165 | 1.6743 | 1.4942 | 0.98677 |
| ” due ” | 161 | 1.3846 | 1.1094 | 0.17369 |
| ” casa ” | 152 | 1.2874 | 1.2264 | 0.80137 |
| ” avevo ” | 151 | 1.4198 | 1.2078 | 0.60146 |
| ” signor ” | 147 | 2.8318 | 1.388 | 0.89798 |
| ” via ” | 147 | 1.1388 | 1.1204 | 0.59772 |
| ” occhi ” | 138 | 1.2546 | 1.0864 | 0.24577 |
| ” sua ” | 137 | 1.3290 | 1.1365 | 0.39867 |
| ” vita ” | 127 | 1.7060 | 1.2852 | 0.74053 |
| ” questa ” | 126 | 1.1847 | 1.0897 | 0.24447 |
| ” papiano ” | 119 | 1.9505 | 1.4917 | 1.7917 |
| ” sul ” | 117 | 1.2827 | 1.2025 | 0.55902 |
| ” mano ” | 110 | 1.2167 | 1.0872 | 0.27922 |
| ” c ” | 110 | 1.0365 | 1.0907 | 0.28188 |
| ” e ” | 2236 | 0.90630 | 1.1206 | 0.54307 |
| ” di ” | 1936 | 1.0455 | 1.1478 | 0.33042 |
| ” che ” | 1861 | 0.93397 | 1.1323 | 0.58236 |
| ” la ” | 1509 | 1.1318 | 1.1734 | 0.52039 |
| ” a ” | 1492 | 1.0240 | 1.1202 | 0.36252 |
| ” non ” | 1356 | 1.0228 | 1.1307 | 0.45989 |
| ” il ” | 1183 | 1.0843 | 1.1858 | 0.47736 |

***La coscienza di Zeno*, N=834656**

| $\alpha$ | $f$ | $B'$ | $\hat{\gamma}$ | $\sigma_{\hat{\gamma}} \times 10^3$ |
|---|---|---|---|---|
| ” guido ” | 569 | 3.1193 | 1.6058 | 1.0319 |
| ” al ” | 566 | 1.0854 | 1.1055 | 0.25708 |
| ” ada ” | 532 | 2.5366 | 1.5518 | 0.78842 |
| ” quella ” | 532 | 1.1571 | 1.1135 | 0.11385 |
| ” augusta ” | 390 | 1.9227 | 1.3848 | 0.40894 |
| ” lo ” | 377 | 1.1479 | 1.0817 | 0.14321 |
| ” tempo ” | 297 | 1.1297 | 1.0497 | 0.14001 |
| ” lui ” | 290 | 1.6458 | 1.2798 | 0.58712 |
| ” carla ” | 280 | 4.5974 | 1.6572 | 0.77453 |
| ” cosi’ ” | 278 | 1.0225 | 1.0441 | 0.23481 |
| ” giorno ” | 276 | 1.2087 | 1.0829 | 0.17055 |
| ” sempre ” | 275 | 1.1532 | 1.0946 | 0.37360 |
| ” casa ” | 262 | 1.4287 | 1.2096 | 0.26115 |
| ” ci ” | 260 | 1.0940 | 1.1157 | 0.26626 |
| ” di ” | 4992 | 0.96690 | 1.0433 | 0.80399 |
| ” che ” | 4232 | 0.91567 | 1.0906 | 0.12420 |
| ” e ” | 3412 | 0.90710 | 1.0923 | 0.089463 |
| ” non ” | 3019 | 0.97300 | 1.0329 | 0.27613 |
| ” la ” | 2904 | 1.0392 | 1.104 | 0.39155 |
| ” il ” | 2319 | 1.0506 | 1.1398 | 0.29722 |
| ” a ” | 2150 | 1.0396 | 1.0849 | 0.33049 |

In the following tabs there are, on the contrary, the words of the same books with highest value of betweenness centrality. Moreover, for these words there are also their values of $B'$ and $\hat{\gamma}$ and an additional column with a ✓if we can consider these words as key-words for both approaches, a $\times$ if we can't and a $\sim$ if it's not so clear.

***Moby Dick***

| Word | $B'$ | $\hat{\gamma}$ | Keyword? |
|---|---|---|---|
| ” whale ” | 2.1503 | 1.555 | ✓ |
| ” man ” | 1.2805 | 1.1944 | × |
| ” ship ” | 1.5666 | 1.3107 | ∼ |
| ” sea ” | 1.3010 | 1.1959 | × |
| ” ahab ” | 3.2174 | 1.5502 | ✓ |

***A naturalist's voyage round the world***

| Word | $B'$ | $\hat{\gamma}$ | Keyword? |
|---|---|---|---|
| ” great ” | 1.1903 | 1.1985 | × |
| ” water ” | 1.5209 | 1.2276 | ∼ |
| ” man ” | 1.5674 | 1.2535 | ∼ |
| ” country ” | 1.5194 | 1.2534 | ∼ |
| ” found ” | 1.2172 | 1.1305 | × |

***Alice's adventures in wonderland***

| Word | $B'$ | $\hat{\gamma}$ | Keyword? |
|---|---|---|---|
| ” alice ” | 0.90337 | 1.0926 | × |
| ” man ” | 1.5141 | 1.2116 | ∼ |
| ” time ” | 0.97476 | 1.1338 | × |
| ”men ” | 1.6074 | 1.1928 | ∼ |
| ” work ” | 0.76962 | 1.0145 | × |

***Life on Mississippi***

| Word | $B'$ | $\hat{\gamma}$ | Keyword? |
|---|---|---|---|
| ” river ” | 2.2193 | 1.4259 | ✓ |
| ” time ” | 1.2064 | 1.0932 | × |
| ” man ” | 1.6388 | 1.2553 | ∼ |
| ” boat ” | 2.0364 | 1.3421 | ✓ |
| ” day ” | 1.1676 | 1.0052 | × |

***The jungle***

| Word | $B'$ | $\hat{\gamma}$ | Keyword? |
|---|---|---|---|
| ” jurgis ” | 2.0933 | 1.4772 | ✓ |
| ” man ” | 1.3175 | 1.2682 | × |
| ” time ” | 1.1973 | 1.1396 | × |
| ” men ” | 1.7881 | 1.3121 | ✓ |
| ” work ” | 1.3733 | 1.2341 | × |

***I Malavoglia***

| Word | $B'$ | $\hat{\gamma}$ | Keyword? |
|---|---|---|---|
| ” ntoni ” | 1.5914 | 1.3578 | ∼ |
| ” casa ” | 1.2793 | 1.1593 | × |
| ” malavoglia ” | 1.2964 | 1.1419 | × |
| ” mena ” | 1.4699 | 1.2912 | ∼ |
| ” andava ” | 1.2325 | 1.0867 | × |
| ” nulla ” | 1.1149 | 1.0911 | × |
| ” fatto ” | 1.0436 | 1.0334 | × |
| ” nonno ” | 1.5388 | 1.3003 | ∼ |
| ” piedipapera ” | 1.5026 | 1.2556 | ∼ |
| ” sempre ” | 1.0576 | 1.0852 | × |

***I pirati della Malesia***

| Word | $B'$ | $\hat{\gamma}$ | Keyword? |
|---|---|---|---|
| ” sandokan ” | 2.5877 | 1.5032 | ✓ |
| ” yanez ” | 2.2748 | 1.3814 | ✓ |
| ” rajah ” | 2.1324 | 1.4204 | ✓ |
| ” kammamuri ” | 1.7348 | 1.2721 | ✓ |
| ” pirati ” | 1.7240 | 1.2298 | ✓ |
| ” tigre ” | 1.9718 | 1.2757 | ✓ |
| ” verso ” | 1.3050 | 1.1427 | × |
| ” capitano ” | 2.9022 | 1.5093 | ✓ |
| ” uomo ” | 1.4047 | 1.1308 | ∼ |
| ” uomini ” | 1.5331 | 1.2380 | ∼ |

***Le avventure di Pinocchio-Storia di un burattino***

| Word | $B'$ | $\hat{\gamma}$ | Keyword? |
|---|---|---|---|
| ” pinocchio ” | 0.90728 | 1.0646 | × |
| ” burattino ” | 1.1020 | 1.0517 | × |
| ” sempre ” | 1.0065 | 1.0999 | × |
| ” dopo ” | 0.97138 | 0.94893 | × |
| ” fatto ” | 0.73374 | 0.88795 | × |
| ” casa ” | 1.5900 | 1.2095 | ∼ |
| ” povero ” | 1.0187 | 1.0250 | × |
| ” ragazzi ” | 1.6898 | 1.3893 | ✓ |
| ” fata ” | 1.9789 | 1.3934 | ✓ |
| ” mai ” | 1.0460 | 0.96949 | × |

***Canne al vento***

| Word | $B'$ | $\hat{\gamma}$ | Keyword? |
|---|---|---|---|
| ” efix ” | 1.7046 | 1.2440 | ✓ |
| ” noemi ” | 2.3610 | 1.4505 | ✓ |
| ” giacinto ” | 2.4431 | 1.3358 | ✓ |
| ” occhi ” | 1.0094 | 1.0178 | × |
| ” casa ” | 1.7921 | 1.1253 | ✓ |
| ” donna ” | 2.4683 | 1.3287 | ✓ |
| ” bene ” | 1.1994 | 1.0246 | × |
| ” pareva ” | 1.0141 | 1.0159 | × |
| ” viso ” | 1.1481 | 1.0964 | × |
| ” sempre ” | 1.2006 | 1.1009 | × |

***Il fu Mattia Pascal***

| Word | $B'$ | $\hat{\gamma}$ | Keyword? |
|---|---|---|---|
| ” adriana ” | 1.6743 | 1.4942 | ✓ |
| ” forse ” | 1.2617 | 1.1177 | × |
| ” casa ” | 1.2874 | 1.2264 | × |
| ” via ” | 1.1388 | 1.1204 | × |
| ” occhi ” | 1.2546 | 1.0864 | × |
| ” vita ” | 1.7060 | 1.2852 | ✓ |
| ” prima ” | 1.3456 | 1.1358 | × |
| ” gia’ ” | 0.95391 | 1.0065 | × |
| ” qualche ” | 1.1372 | 1.1074 | × |
| ” fatto ” | 1.3742 | 1.0644 | × |

***La coscienza di Zeno***

| Word | $B'$ | $\hat{\gamma}$ | Keyword? |
|---|---|---|---|
| ” guido ” | 3.1193 | 1.6058 | ✓ |
| ” ada ” | 2.5366 | 1.5518 | ✓ |
| ” essa ” | 1.9040 | 1.3867 | ✓ |
| ” augusta ” | 1.9227 | 1.3848 | ✓ |
| ” prima ” | 1.0390 | 1.0590 | × |
| ” sempre ” | 1.1532 | 1.0946 | × |
| ” carla ” | 4.5974 | 1.6572 | ✓ |
| ” qualche ” | 1.0551 | 1.0266 | × |
| ” grande ” | 1.0724 | 1.0763 | × |
| ” giorno ” | 1.2087 | 1.0829 | × |

In order to better understand these results, in the next figures there are Burstiness-Correlation diagrams for those words extracted from *Moby Dick* and from *La coscienza di Zeno*, using Filippo's and Giulia's method.

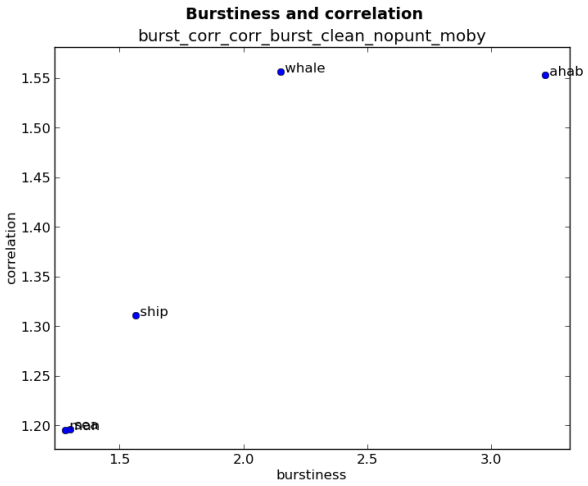



Figure 5.1: Burstiness-Correlation diagram for the key-words extracted from *Moby Dick*.

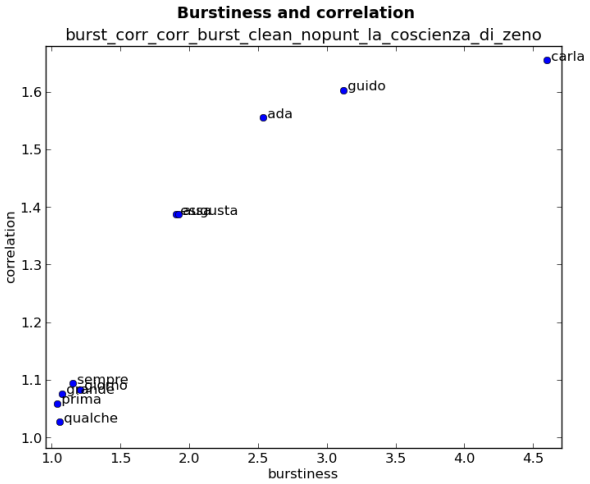



Figure 5.2: Burstiness-Correlation diagram for the key-words extracted from *La coscienza di Zeno*.

As we can see from these results, the maximum analogies are obtained for the longest novels in both languages and the main reason of this result is probaly the fact that my approach studies asymptotic behaviors and results obtained from this approach is probably as more precise, as longer is the text used. This may let us argue that these two approaches may lead to similar results for long "enough" texts, but this fact should obviously be tested and deepened.

# Chapter 6

# Conclusions

The idea of working with different translations of the same book was born in order to check and confirm the hypothesis, proposed by E. G. Altmann, G. Cristadoro and M. Degli Esposti in [43] and explained in the third chapter. They argue that, for any condition $\alpha$, its associated random walker $X(t)$ will spread super-diffusively as $\sigma_X^2(t) \propto t^\gamma$ with the same exponent $\gamma$. This means that, asymptotically, $\gamma$ should be equal for every chosen $\alpha$. Since different translations of the same book can be considered as a particular "shuffling" procedure which fixes topics at the topic level in the hierarchy of language, my approach to analyze differences and analogies between key-words' behaviors had exactly this goal. Thus, the most important result to show is certainly the fact that the choice of different languages don't influence neither the presence, nor the exponent of the long-range correlations. In fact, while the most frequent words, that may have many possible translations between different languages (for example the word *" the "* in English can be translated in Italian with " il ", " la ", " lo ", etcetera) and whose use may strongly depend on particular grammar rules, exhibit similar but consistently different evaluates of the exponents of long-range correlations between different languages, keywords, that are used only in particular topics (for example the word *" prince "* in *War and Peace* is used only when Tolstoj is talking about war), exhibit evaluates of the exponents of long-range correlations that are

almost equal for every language used. This result, obtained using this particular "shuffling" method, strongly confirms the hypothesis that, in a text, the topic is really at the highest level of our hierarchy and it is what actually characterizes long-range correlations exponent.

Another important result obtained in this thesis is that, although almost every letter and most frequent words exhibit a value of $B' \approx 1$ and may thus be imagined, from an approximate point of view, as results of a Poisson process, experimental results, explained in the fourth chapter, highlights that all these sequences are long-range correlated without burstiness and so their non-Poissonian nature, and thus their information richness, is revealed through long-range correlations, $\gamma > 1$.

# Appendix A

# Texts cleaning

For my analysis I obviously couldn't use texts as can be downloaded (the books used for my experiments was downloaded from [48], [49] and [50]), so I decided to clean the texts as explained in the following tab:

| Original character | Used character | | Original character | Used character |
|---|---|---|---|---|
| "A" | "a" | | "B" | "b" |
| "C" | "c" | | "D" | "d" |
| "E" | "e" | | "F" | "f" |
| "G" | "g" | | "H" | "h" |
| "I" | "i" | | "J" | "j" |
| "K" | "k" | | "L" | "l" |
| "M" | "m" | | "N" | "n" |
| "O" | "o" | | "P" | "p" |
| "Q" | "q" | | "R" | "r" |
| "S" | "s" | | "T" | "t" |
| "U" | "u" | | "V" | "v" |
| "W" | "w" | | "X" | "x" |
| "Y" | "y" | | "Z" | "z" |

| Original character | Used character | | Original character | Used character |
|---|---|---|---|---|
| ”,” | ” ” | | ”;” | ” ” |
| ”.” | ” ” | | ”:” | ” ” |
| ”-” | ” ” | | ”_” | ” ” |
| ”#” | ” ” | | ”””” | ” ” |
| ”[” | ” ” | | ”]” | ” ” |
| ”(” | ” ” | | ”)” | ” ” |
| ”?” | ” ” | | ”!” | ” ” |
| ”/” | ” ” | | ”*” | ” ” |
| ”=” | ” ” | | ”’” | ” ” |
| ”¡” | ” ” | | ”¿” | ” ” |
| ”\t”(Tab) | ” ” | | ”\n”(Newline) | ” ” |
| ”ê” | ”e” | | ”à” | ”a”” |
| ”é” | ”e”” | | ”ó” | ”o”” |
| ”ä” | ”a” | | ”è” | ”e”” |
| ”ù” | ”u”” | | ”ì” | ”i”” |
| ”ë” | ”e” | | ”û” | ”u” |
| ”ô” | ”o” | | ”$\overline{\text{o}}$” | ”o” |
| ”ò” | ”o”” | | ”â | ”a” |
| ”î” | ”i” | | ”ö” | ”o” |
| ”ü” | ”u” | | ”$\bar{\text{i}}$” | ”i” |
| ”õ” | ”o” | | ”◦” | ” ” |
| ”Ü” | ”u” | | ”$\beta$” | ”ss” |
| ”À” | ”a”” | | ”ú” | ”u”” |
| ”Ó” | ”o”” | | ”Ä” | ”a” |
| ”Ö” | ”o” | | ”Ò” | ”o”” |
| ”É” | ”e”” | | ”È” | ”e”” |
| ”Ê” | ”e” | | ”Ç” | ”ç” |
| ”Ô” | ”o” | | ”$” | ” ” |
| ”æ” | ”ae” | | ”Ù” | ”u” |

# Bibliography


[1] G. K. Zipf, *The psycho-biology of language*. Oxford, England: Houghton, Mifflin (1935)

[2] G. K. Zipf, *Human behavior and the principle of least effort*. Oxford, England: Addison-Wesley Press (1949)

[3] H. A. Simon, *On a class of skew distribution functions*. Biometrika (1955)

[4] G. A. Miller, *Some Effects of Intermittent Silence*. The American Journal of Psychology (1957)

[5] B. Mandelbrot, *The Pareto-Levy Law and the Distribution of Income*. International Economic Review (1960)

[6] G. Herdan, *Quantitative linguistics*. London: Butterworths (1964)

[7] H. S. Heaps, *Information retrieval: computational and theoretical aspects*. Academic Press, Inc. Orlando, FL, USA (1978)

[8] G. Kirby, *Zipf's law*. UK Journal of Naval Science (1985)

[9] S. Naranan and V. K. Balasubrahmanyan,*Information theoretic models in statistical linguistics- Part I: A model for word frequencies*. Current science (1992)

[10] S. Naranan and V. K. Balasubrahmanyan,*Information theoretic models in statistical linguistics- Part II: Word frequencies and hierarchical structure in language.statistical tests*. Current science (1992)

[11] R. F. Voss, *Evolution of Long-Range Fractal Correlations and $\frac{1}{f}$ Noise in DNA Base Sequences*. Physical Review Letters (1992)

[12] W. Li, *Random Texts Exhibit Zipf's-Law-Like Word Frequency Distribution*. IEEE Transactions on Information Theory (1992)

[13] G. Trefan, E. Floriani, B. J. West and P. Grigolini, *Dynamical approach to anomalous diffusion: Response of Levy processes to a perturbation*. Physical Reviw E (1994)

[14] W. Ebeling and T.Pöschel, *Entropy and Long range correlations in literary English*. Europhysics Letters (1994)

[15] M. Amit, Y. Shmerler, E. Eisenberg, M. Abraham and N. Shnerb, *Language and Codification Dependence of Long-Range Correlations in Texts*. Fractals (1994)

[16] W. Ebeling and A. Neiman, *Long-range correlations between letters and sentences in texts*. Physica A (1994)

[17] R. Baeza-Yates and G. Navarro, *Block Addressing Indices for Approximate Text Retrieval*. JASIS (1997)

[18] G. Rangarajan and M. Ding, *Integrated approach to the assessment of long range correlation in time series data*. Phys Rev E (2000)

[19] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Massachusetts Institute of Technology (2000)

[20] R. F. i Cancho and R. V. Solè, *Two Regimes in the Frequency of Words and the Origins of Complex Lexicons: Zipf's Law Revisited*. Journal of Quantitative Linguistics (2001)

[21] M. A. Montemurro, *Beyond the Zipf-Mandelbrot law in quantitative linguistics*. Physica A: Statistical Mechanics and its Applications (2001)

[22] A. Gelbukh and G. Sidorov, *Zipf and Heaps Laws' Coefficients Depend on Language*. Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing (2001)

[23] L. A. Adamic and B. A. Huberman, *Zipf's law and the Internet*. Glottometrics (2002)

[24] M. Ortuño, P. Carpena, P. Bernaola-Galván, E. Muñoz and A. M. Somoza, *Keyword detection in natural languages and DNA*. Europhysics Letters (2002)

[25] M. A. Montemurro and P. A. Pury, *Long-range fractal correlations in literary corpora*. Fractals (2002)

[26] R. F. i Cancho and R. V. Solè, *Zipf's law and random texts*. Advances in Complex Systems (2002)

[27] R. F. i Cancho and R. V. Solè, *Least effort and the origins of scaling in human language*. PNAS (2003)

[28] P. Allegrini, P. Grigolini, and L. Palatella, *Intermittency and Scale-Free Networks: A Dynamical Model for Human Language Complexity*. Chaos, Solitons and Fractals (2004)

[29] D.C. van Leijenhorst and Th.P. van der Weide, *A formal derivation of Heaps' Law*. Information Sciences (2005)

[30] M. E. J. Newman, *Power laws, Pareto distributions and Zipf's law*. Contemporary Physics (2005)

[31] D. H. Zanette and M. A. Montemurro, *Dynamics of text generation with realistic Zipf distribution*. Journal of Quantitative Linguistics (2005)

[32] T. M. Cover and J. A. Thomas,*Elements of Information Theory-Second Edition*. Wiley-Interscience, Hoboken (2006)

[33] Benjamin Lindner, *Superposition of many independent spike trains is generally not a Poisson process*. Physical Review E (2006)

[34] A. Vázquez, J. G. Oliveira, Z. Dezsö, K.-I. Goh, I. Kondor, and A. L. Barabási1, *Modeling bursts and heavy tails in human dynamics*. Physical Review E (2006)

[35] L. da F. Costa, F. A. Rodrigues, G. Travieso, P. R. Villas Boas, *Characterization of Complex Networks: A Survey of measurements*. Advances in Physics (2007)

[36] K. I. Goh and A. L. Barabasi, *Burstiness and memory in complex systems*. EPL (Europhysics Letters) (2008)

[37] S. Naranan and V. K. Balasubrahmanyan, *Models for power law relations in linguistics and information science*. Journal of Quantitative Linguistics (2008)

[38] E. G. Altmann, J. B. Pierrehumbert and A. E. Motter, *Beyond Word Frequency: Bursts, Lulls, and Scaling in the Temporal Distributions of Words*. PLOS one (2009)

[39] L. Lü, Z. K. Zhang and T. Zhou, *Zipf's Law Leads to Heaps' Law: Analyzing Their Relation in Finite-Size Systems*. PLOS one (2010)

[40] R. F. i Cancho and B. Elevåg, *Random Texts Do Not Exhibit the Real Zipf's Law-Like Rank Distribution*. PLOS one (2010)

[41] S. Bernhardsson, S. K. Baek and P. Minnhagen, *A Paradoxical Property of the Monkey Book*. Journal of Statistical Mechanics: Theory and Experiment (2011)

[42] D. H. Zanette, *Statistical Patterns in Written Language*. (2012)

[43] E. G. Altmann, G. Cristadoro and M. Degli Esposti *On the origin of long-range correlations in texts*. PNAS (2012)

[44] A. Baronchelli, V. Loreto and F. Tria, *Language Dynamics*. Advances in Complex Systems (2012)

[45] M. Gerlach and E. G. Altmann, *Stochastic model for the vocabulary growth in natural languages*. American Physical Society (2013)

[46] E. G. Altmann, Z. L. Whichard, A. E. Motter,*Identifying Trends in Word Frequency Dynamics*. Journal of Statistical Physics (2013)

[47] M. A. Montemurro, *Quantifying the information in the long-range order of words: semantic structures and universal linguistic constraints*. Cortex (2013)

[48] http://www.gutenberg.org

[49] http://gutenberg.spiegel.de

[50] http://LiberLiber.it

# Ringraziamenti

Non sono il tipo da scrivere queste cose melense, per cui proveró a non esserlo.

Ringrazio innanzitutto i Professori Mirko Degli Esposti e Giampaolo Cristadoro per la loro disponibilitá e la costanza con cui mi hanno seguito, per aver condiviso le loro idee con me e per avermi consigliato la *IQLA-GIAT Summer School in Quantitative Analysis of Textual Data*. Ne approfitto per ringraziare tutti coloro che hanno organizzato e partecipato a questa Summer School, fonte di idee che hanno contribuito non poco nel mio percorso per giungere a questa tesi. Ringrazio inoltre Filippo e Giulia, con i quali ho lavorato in quest'ultimo periodo.

Ringrazio la mia famiglia: i miei genitori che mi hanno sempre supportato, ma soprattutto sopportato, e mia sorella, che ovunque sia e ovunque vada, c'é sempre.

Ringrazio la Giada, punto fisso di questi anni universitari (e si spera oltre), per tutto ció che ha fatto per e con me.

Ringrazio la Metrey per insegnarmi ogni volta qualcosa con la sua spontanea e genuina fratellanza e la tata, perché é e rimane la "tata zia".

Ringrazio i miei nonni perché é anche grazie al loro sudore e al loro senso del lavoro se sono arrivato fino a qui.

Ringrazio i miei amici storici per tutte le cose condivise insieme, dalle birre in Lunetta, alle vacanze in Albania.

Ringrazio i miei amici di facoltá, non solo per aver condiviso con me questo percorso, ma soprattutto perché mi sono veramente amici. Alcuni purtroppo, governo ladro, non possono essere qua a condividere questo momento con me, ma so che ad Aarhus e Philadelphia gli fischieranno le orecchie.

Ringrazio i miei compagni della BaLotta Continua per quest'ultimo meraviglioso anno passato insieme a suonare in giro per l'Italia (piú o meno...).

Ringrazio le Professoresse Ileana Civili e Carla Rossi per essere state coloro che per prime mi hanno fatto appassionare alla matematica.

Ringrazio tutti coloro che, in un modo o in un altro, mi hanno fatto diventare ció che sono e mi hanno fatto giungere a questo traguardo.