Universidade
Federal de
Minas Gerais

Statistical Analyses in Language Usage

Leonardo Carneiro de Araújo

Orientador: Hani Camille Yehia

Trabalho apresentado ao Programa de ao Programa de Pós-graduação em Engenharia Elétrica da Escola de Engenharia da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do grau de Doutor em Engenharia Elétrica.

Outubro 2013

Programa de Pós-Graduação em Engenharia Elétrica
Universidade Federal de Minas Gerais

Abstract

Language has a fundamental social function, it is a widely used mean of communication, dynamic, robust and still so simple; a specific human capacity, capable of carrying our thoughts and maybe the only feature that make us humans fundamentally different from other species, and still so vaguely understood. Approximately from 3000 to 7000 languages are spoken nowadays, all of them hold remarkable distinctions one from another, but still have much in common. Recent research on cognitive sciences has concluded that patterns of use strongly affect how language is perceived, acquired, used and changes over time. It is argued that languages are self-organizing systems, and that language usage creates and shapes what languages are. The linguistic competence of a speaker is attributed to self-organization phenomena, but not to a nativist hypothesis. The purpose of this study is to develop statistical analyses of language usage based on a detailed investigation of the Zipf's law and other laws of quantitative linguistics. We will develop a systematic empirical investigation of phenomena via statistical, mathematical and computational techniques. We carry out, first, a horizontal analysis across different languages using the UCLA Phonological Segment Inventory Database. This analysis is followed by a vertical investigation of English patterns in different linguistic structural levels. In addition to the results obtained with Zipf's law, information theoretical analyses are done in order to understand the trade-off between the efficiency of language information transmission and language complexity. We observe that the features of linguistic elements and their interrelations abide by universal laws (in the stochastic sense). These analyses are important for a quantitative comprehension of linguistic concepts that are already well known qualitatively, providing a means to understand the processes underlying language usage and evolution. Understanding how languages works and evolves might be the only hope to create technological artifacts that truly exhibit human-level communication capabilities, being able to understand and produce human-like sentences/utterances.

Resumo

A linguagem possui uma função social fundamental, ela é uma forma de comunicação amplamente utilizada, dinâmica, robusta e ainda assim tão simples; uma faculdade específica dos humanos, capaz de levar nossos pensamentos e talvez a única característica que nos distinga de outras espécies; e ainda tão pouco compreendida. Aproximadamente de 3000 a 7000 línguas são faladas nos dias atuais, todas possuem diferenças marcantes em relação às outras, entretanto possuem muito em comum. Pesquisas recentes em ciências cognitivas demonstraram que os padrões de uso influenciam fortemente a maneira como a linguagem é percebida, adquirida, utilizada e como ela muda ao longo do tempo. Defende-se que as línguas são sistemas auto organizativos, e que o próprio uso da linguagem cria e molda o que elas são. Atribui-se a competência linguística de um falante a um fenômeno auto organizativo, ao invés de uma hipótese inata. O propósito deste estudo é desenvolver uma análise estatística do uso da língua a partir da investigação minuciosa da lei de Zipf e outras leis de linguística quantitativa. Iremos desenvolver uma abordagem empírica sistemática de investigação dos fenômenos através de técnicas estatísticas, matemáticas e computacionais. Primeiramente faremos uma análise horizontal ao longo de diferentes línguas utilizando o banco de dados de inventários fonológicos segmentais criado pela UCLA. Esta análise será seguida por uma análise vertical investigando os padrões do Inglês em diferentes níveis estruturais linguísticos. Além dos resultados obtidos para a lei de Zipf, uma análise sob a ótica da teoria da informação é feita para entender a relação de compromisso entre eficiência em transmissão de informação de uma língua e complexidade da linguagem. Observamos que as propriedades dos elementos linguísticos e suas inter-relações seguem leis universais (no sentido estocástico). Estas análise são importantes para a compreensão quantitativa de conceitos linguísticos que são bem conhecidos de forma qualitativa, fornecendo assim os meios para entender o uso da língua e sua evolução. Entender como funcionam as línguas e como elas evoluem pode ser a única maneira de se criar artefatos tecnológicos que realmente possuem uma capacidade de comunicação equiparável à humana, sendo assim capaz de entender e produzir sentenças/elocuções semelhantes àquelas produzidas pelos homens.

Contents

List of Figures

List of Tables

1 Introduction

Language is a biological, psychological and social process. The study of language as a communication process involves insight on these subjects and a scientific analysis of data produced as a mean of information transfer. Performing a statistical analysis of language is a way of acknowledging its unpredictable nature, as the uncertainty intrinsic to it is the way in which it is possible to carry information. Although language has a random nature, it holds an order, coordination and structuration that imposes an amount of redundancy to the transmitted message. It is important to characterize the process and understand what variables are into play in the communication process. Language is not a process controlled by a single agent, rather it is driven by interactions of multiple agents, it is wholly decentralized or distributed over all the components of the system.

All languages attain such characteristics and therefore it is important to analyze languages from this common ground and try to understand, based on the common patterns observed in languages, how languages work. We need then to change our paradigm of 'linguistic universals'. As we might observe, language speech inventories are quite diverse and there are vanishingly few linguistic universals in direct sense left. On the other hand, as we regard language as a adaptive complex system, we shall observe that there are patterns in language that are also usual in natural phenomena. The ubiquity of power laws is the most notorious one and for that reason we will deeply investigate the well know Zipf's law.

In this study the focus will be on a statistical analyses of language usage, performing a detailed investigation of quantitative linguistics laws. The approach chosen consist on first analyze different languages and then perform a deep examination on English patterns. The communication process is observed under the information theoretical point of view to understand the relation existing between the efficiency in information transfer and the complexity of the system. These analyses are important to comprehend how the language communication phenomenon works and correlate the findings with the well known linguistic concepts.

The analysis of language as a complex system is radically different from the traditional analysis based on a static system of grammatical principles, as a result of the generativist approach. This new approach to language is important for it may allow a unified understanding of seemingly unrelated linguistic phenomena, such as: "variation at all levels of linguistic organization; the probabilistic nature of linguistic behavior; continuous change within agents and across speech communities; the emergence of grammatical regularities from the interaction of agents in language use; and stagelike transitions due to underlying nonlinear processes"(Beckner et al., 2009).

The language patterns are important for language usage, acquisition and efficiency. A well known example is the word frequency effect on lexical access (Whaley, 1978; Grainger, 1990; Andrews, 1989). Low frequency words require greater effort than high frequency words on recognition task, leading to a poorer performance on speed and accuracy tests. Words might be ranked in order of their frequencies of occurrence and that leads to the observation of a power law relation between word rank and frequency. Length of words is also not a mere hazard but a rational deliberation aiming a thrifty and efficient use of resources in a communication process. The way a language sound system is organized seeks a maximal dissimilarity between stimuli. This is an important choice in order to convey maximal information transfer between speaker and listener in a noisy environment. In this huge universe of multiple possible combination of structures, we believe the formation of a language is guided by choices, which organize and structure the random process of communication. Languages are complex systems whose emergence is an event of central importance to human evolution. Several remarkable features suggest the presence of a fundamental principle of organization that seems to be common among all languages.

In this dichotomy of 'language as chance' - 'language as choice', applying quantitative methods is fundamental to let us draw insights on nature of this communication phenomenon. This dichotomy, rightly understood, might appear as the bridge between the two dichotomies proposed by Saussure: 'langue-parole' and 'significant-signifié. "In fact, the relation is quite close: language as change refers to the langue-parole dichotomy in its interpretation as that between statistical universe and sample, whereas language as

choice refers to the signifiant-signifié dichotomy in its interpretation as being subject to the law of duality"(Herdan, 1966).

"If a statistical test cannot distinguish rational from random behavior, clearly it cannot be used to prove that the behavior is rational. But, conversely, neither can it be used to prove that the behavior is random. The argument marches neither forward nor backward"(Miller, 1965). Contrary to Miller's belief, we argue that a statistical characterization of language as a communication process is of central importance to trace the line that distinguish a mere random event from another, also random in nature, but that stands in the watershed between chaos and order, establishing a balance between information transfer and communication cost.

The idea of statistical treatment of language data is not new, and we might even say that linguistics is not possible without some degree of statistical classification. Linguists have always used patient recording, annotations and classifications in order to imagine what would be a possible grammar for that language. Moreover, a regularity in the historic observation of language, like the Grimm's law consonantal shift, could only be realized after an investigation on a long and patient collection of data. Comparative philology also uses the comparison of a great mass of linguistic data to establish the relationships among languages and families.

"The effectiveness of language as a means of communication depends, naturally, on its being highly patterned, and hence on its users' behaviour being predictable, not necessarily as to the meanings they will convey in each individual situation, but as to the phonological, morphological, and syntactical paths they will follow in so doing. Yet no set of speech-habits is entirely rigid or ultrasystematic... There are always loose ends within the system of speech behaviour. It is this inherent looseness of linguistic patterning, together with built-in redundancy, that makes change not only normal but inevitable, and thus a basic part of language.

The great mistake of the idealists (determinists) is their overemphasis on vocabulary choice as the only source of linguistic change, and their consequent neglect of the habitual aspects of language. Our linguistic behaviour is very largely a matter of habit, and, in Twadell's words, 'below and above the control of the individual' – below because it is so largely unreflecting habit in brain, nerve, and muscle; above, because it is so largely influenced, from its very inception in each of us, by the behaviour of other members of the community.

Each individual builds up his own set of speech-habits, his idiolect, in himself, and of course the idiolect is the only ultimate linguistic reality. Entities such as 'dialect' or 'languages' are always abstractions formed on the basis of a comparison of two or more idiolects... Yet this does not mean that each individual 'creates' his language ex novo;

virtually all our speech-habits are built up through imitation of those of other individuals, and what little is 'original' with each speaker derives from combination of already existing patterns. An idiolect is effective as a means of communication only because it closely resembles the idiolects of other speakers. There is never an absolute identity between any two idiolects, but there can be a very close similarity which justifies our abstracting (naively or analytically) what is common to them and treating it as an entity.

Each language, each dialect has its phonemic structure, and only what is within that structure is possible for the speaker and listener of the language or dialect. And within the limits of structure imposed by the community, the individual speaker makes his choices... He sees his choices as free and... comes to ignore the limitations and move about them comfortably, so that the real choices become the only choices he sees"(Hall, 1964).

In order to capture language as an emergent identity on the vast universe of idiolects and spoken realizations, it is important to observe the recurring patterns on a large dataset and extract linguistic meaning from it. The quantitative analysis of languages is important to produce a systematic empirical investigation of the language phenomenon via statistical, mathematical or computational techniques. It is grounded on a large data of empirical observations, which are used to develop and employ mathematical models, theories and hypothesis pertaining the phenomenon.

The quantitative approach to language analysis data back to the ancient Greek who have used combinatorics to investigate the formation of linguistic structures. Later, the philologist and lexicographer Al-Khalil ibn Ahmad (718-791) used permutations and combinations to list all possible Arabic words with and without vowels. William Bathe (1564-1614) published the word's first language teaching texts, called 'Janua Linguarum', where he had compiled a list with 5.300 essential words, according to their usage. From the end of the 19th century many scientific work on language started using the quantitative approach. Augustus De Morgan (1851), for example, on the statistical analysis of literary style, suggested that one could identify an author by the average length of his words. Many scientific counts of units of language or text were published in the 19th century as a means of linguistic description: in Germany, Förstemann (1846, 1852) and Drobisch (1866); in Russia, Bunjakovskij (1847); in France, Bourdon (1892); in Italy, Mariotti (1880); and in the USA, Sherman (1888). From the 20th century on, many scientific work has been produced on quantitative linguistics. Among many names we might cite: Andrey Andreyevich Markov, George Kingley Zipf, Benoît Mandelbrot, Claude Elwood Shannon and Warren Weaver, Gustav Herdan, Rajmund.G. Piotrowski, Walter Meyer-Eppler, Gabriel Altmann, Reinhard Köhler, Paul Menzerath, Juhan Tuldava, Peter Grzybek and Wilhelm Fucks.

The quantitative approach in linguistics gained a big boost with the work of George

Kingley Zipf. The idea proposed by Zipf (1949) is that language works seeking the principle of least effort. This theory is also applied to different fields such as evolutionary biology. It postulates that it is a natural behavior to seek a path of low cost. Although this principle seems reasonable to be applied to all languages and their evolution, it creates an opposition with other features needed so that a communication processes may be held, such as variability and distinguishability. This trade-off might be seen as a result of the conflicting application of the principle of least effort to the speaker and the hearer. Speakers want to minimize articulatory effort, use sequences of phones that are easy to pronounce and encourage brevity and phonological reductions. On the other hand, hearers want to minimize the decoding effort of uttered speech, it is necessary then to enhance explicitness, clarity and distinguishability. The hearer wants to avoid ambiguity in the lexical level, minimizing the effort to understand a sentence. The speaker will tend to choose the most frequent words, which tend to be the most ambiguous ones (Gernsbacher, 1994; Köhler, 1986). Zipf (1949) pointed out that this lexical trade-off could explain the pattern of word frequencies, although no rigorous proof of its validity was ever given.

"Underlying the endless and fascinating idiosyncrasies of the world's languages there are uniformities of universal scope. Amid infinite diversity, all languages are, as it were, cut from the same pattern. Some interlinguistic similarities and identities have been formalized, others not, but working linguists are in many cases aware of them in some sense and use them as guides in their analyses of new languages" (Greenberg et al., 1966).

All languages exhibit two distinguished traits: syntax and symbolic reference (Chomsky, 1968; Deacon, 1997). As they are always used as a communication system, that uses the same physical medium to transport information, uses the same biological apparatus to encode and decode the transmitted messages and is a mean of social interaction, being shared among a community, they should also share other characteristics regarding their constituents parts, structure and usage.

The linguistic analysis of a language is the observation of certain recurring patterns, their transformation over time and interactions. Patterns that occur systematically across natural languages are called linguistic universals. An important goal of linguistics is to explain the reason why these patterns emerge so often, which is also a concern of cognitive studies. Some approaches might be used to carry out a systematic research and to analyze the role of these regularities on languages. We are here concerned with a statistical analysis based on real world data, through the usage of linguistic corpora, and with computer simulations of models mimicking language interactions.

We know that speech sounds used in spoken communication vary from one language to the other. We propose to perform a statical analysis of the speech inventories used in different languages. For this purpose we will use the UCLA Phonological Segment

Inventory Database (UPSID) which has 451 languages in its database. We will observe the different speech inventories used and their characteristics. Among these various languages, we will observe that some speech sounds are very common while others are quite rare. All these analyses presupposes that a speech utterance might be segmented into distinctive speech segments, phones. The UPSID has a detailed description of the phones used in each language and much information might be extracted by means of this database.

It is still unclear what is the nature of the language constituent elements, how they are used and organized, and how they change over time. The phoneme, taken as a mental representation, the basic element of spoken language, has been questioned over its status on the study of language (Port, 2007; Port and Leary, 2005; Port, 2006). Port (2007) argues that "words are not stored in memory in a way that resembles the abstract, phonological code used by alphabetical orthographies or by linguistic analysis". According to him, the linguistic memory works as an exemplar memory, where the information stored is an amalgam of auditory codes which include nonlinguistic information. The acceptance and usage of the phonetic model is a reflex of our literacy education (Port, 2007; Coleman, 2002). The assumption of a segmental description of speech is also desired since it guarantees a discrete description at the lower level, what implies discreteness at all other levels. All formal linguistics is based on one *a priori* alphabet of discrete tokens.

Some results, pointed by Port (2007), are against this segmental view of language. The familiarity with one speaker's voice improves the speech recognition at approximately 6%, and this improvement is increased slightly as the variability of the others speakers increase. Port (2007) also argues that richness in dialect variation and language change might not be explained when language information is not stored in a detailed form. Another argument is the well-known frequency effect. When listening to words in noise, the most frequent words in the language can be more accurately recognized than less frequent words. It is also known that "the frequency of words and phrases can have a major influence on speech production in most languages. Typically frequent words suffer greater lenition, that is, reduction in articulatory and auditory distinctness, than infrequent words"(Port, 2007).

The idea of discrete entities being born from the continuous is an interesting one, for correspondence can be drawn with the information carried by a continuous energy process. A discrete system is assumed to involve higher levels of organization. It was always obvious that spoken language has a continuous substratum, but it was the major objective of linguistic structuralism to describe phonology as a system built on discrete entities and logical rules, what would impose a discrete structure on a phonetic continuum (Chomsky, 1957). Mandelbrot (1954) argues that, at the phonological level of language, discreteness is a necessary feature and there is a necessary relationship between continuity

and discreteness in linguistic change. Wilden (2001) points out that "digitalization is always necessary when certain boundaries are to be crossed, boundaries between systems of different *types* or of different *states*, although how these types or boundaries might be operationally defined is unclear".

We will propose here one analysis which consists of using large text databases as our language corpora. We intend to analyze the data at different levels and for this purpose we use pronouncing dictionaries (to perform analysis at the phonemic level), and syllabification dictionary transcription (to analyze syllables). We are then able to procure statistical information on the usage of phones, diphone, triphones, syllables and words in a language. Although written and spoken languages present some marked differences, we assume it is still reasonable to estimate the phonological patterns in a spoken language through the patterns observed in its written counterpart, when a transcription into the phonemic-level is used. Spoken language tend to be more ragged, repetitions are more frequent and the vocabulary is smaller. Even if these differences exist, we assume that the patterns in a spoken language might be, at least coarsely, estimated using written texts. The focus of our analysis will be on the English language, for that reason we used English text databases and an English Pronouncing Dictionary. Each text word was transcribed into phonemic writing using the Carnegie Mellon University Pronouncing Dictionary. Although the words appear in the text within a context, the interactions between the last sounds of a previous word and the initial sounds of the following word were neglected. The analysis of sound structures was restricted to words structures and the text database was used only to acquire a statistical estimation on the frequency in which English phones occur. The same analysis here proposed may be applied to other languages, given a text database and a pronouncing dictionary on that language. If languages are organized through a similar approach, it might be possible to find recurrent patterns as we analyze different languages.

Although "languages are simultaneously products of history and entities existing at particular times"(Good, 2008), both diachronic and synchronic aspects are important to determine what languages are. We focus here on the synchronic aspects. The diachronic approach is also important to investigate since it might clarify how languages change and even determine how usage is responsible for these changes. Language is a social construct and so it is driven by human society.

The statistical analysis of language organization may be essential to determine what sound patterns tend to be ubiquitous and what are the linguistic universal. As pointed out by Mielke (2005), "phonetically coherent classes like /m n ŋ/ and /u o ɔ/ seem to recur in different languages, while more arbitrary groupings like /m n tʃ/ and /ɪ ʔ k^w/ are less common". What might be explained, according to Mielke (2005) by two different

claims: (1) an innatist claim, that argues that common classes (sounds patterns) may be described by a conjunction of distinctive features; (2) an emergentist claim, in which common classes are the result of a common fate. It is also important to understand how these classes are built and used within a language. This sort of analysis might be useful to address issues like that.

One important aspect of understanding language organization and evolution is to understand the interrelations among the different phones in a language. The contrasts we make between sounds are an important aspect to define phonetic similarity, which "is a prerequisite for using it to account for phonological observations"(Mielke, 2005). It is still not clear which might be the grounds to establish such a metric, but it ought to be investigated through a statistical analysis of how speech systems are organized and used. Mielke (2005) pointed that "what is needed is a similarity metric based on objective measurements of sounds. In order to develop such a metric, it is necessary to choose sounds to measure and ways to measure them".

We will use the distinctive features to propose a dissimilarity measure between speech segments. The distinctive features are a combination of articulatory and perceptual features. They are expressed in a binary vector and the features are such that every two speech sounds differ by at least one feature. The proposed dissimilarity measure establishes a separation among phones in a way that resembles the notion of *natural class*, a set speech sounds that share certain phonetic features and undergo the same phonological rules.

In the present work we shall also analyze some statistical features of the English language, the occurrence of patterns and the information content transmitted in a message. Some recurring behaviours are referred to as *laws*, in the stochastic sense, and they are described mathematically in the context of quantitative linguistics. Those mathematical descriptions, along with statistics and information theory, are used to model language characteristics and inquire on its structures, usage and evolution.

2

Language

> But the most noble and profitable invention of all others was that of speech ... whereby men register their thoughts, recall them when they are past, and also declare them to one another for mutual utility and conversation; without which there had been amongst men neither commonwealth, nor society, nor contract, nor peace, no more than amongst lions, bears and wolves.
>
> Thomas Hobbes (1651), Leviathan.

Language refers to the forms of communication among people, or a human capacity for acquiring and using a complex system of communication. Various means might be used to achieve communication, such as gesture, facial expressions, written text, etc. The most usual is the spoken realization of language, which will be our main concern here. Since centuries ago, the human faculty of language has intrigued and motivated the studies seeking to understand what is a language and how it works. All efforts made so far have brought until these days only a fragmentary and superficial insight of the communication phenomena. It is important to realize that language is not just a

collection of words that happen to occur in succession, it is the grammar that express how we put the words together in order to convey meaning. Although not yet understood, this process is so simple that nearly every child masters it almost unconsciously. The complex set of rules and patterns that creates spoken communication is mastered by any individual. "A grammar of a language purports to be a description of the ideal speaker-hearer's intrinsic competence" (Chomsky, 1969) and still it is not fully understood by the language scientists.

Language is intrinsically bound to thought. Our thoughts are stated under our language knowledge, and we use it as a tool to express them. Our linguistic reality, the categories and usage in language, are said to shape how we perceive the world and the way we think. That is called 'the Sapir-Whorf hypothesis', named after the American linguists Edward Sapir and Benjamin Lee Whorf.

> Human beings do not live in the objective world alone, nor alone in the world of social activity as ordinarily understood, but are very much at the mercy of the particular language which has become the medium of expression for their society. It is quite an illusion to imagine that one adjusts to reality essentially without the use of language and that language is merely an incidental means of solving specific problems of communication or reflection. The fact of the matter is that the 'real world' is to a large extent unconsciously built upon the language habits of the group. No two languages are ever sufficiently similar to be considered as representing the same social reality. The worlds in which different societies live are distinct worlds, not merely the same world with different labels attached... We see and hear and otherwise experience very largely as we do because the language habits of our community predispose certain choices of interpretation. (Sapir, 1929)

Comparing languages, it is possible to notice that the categorization of the world is different across cultures. Whorf (1956), in a popular paper, refereed to Eskimo languages as having distinct categorizations for types of snow and therefore different words were used. "We find that the idea of *water* is expressed in a great variety of forms: one term serves to express water as a *liquid*; another one, water in the form of a large expanse (*lake*); others, water as running in a large body or in a small body (*river* and *brook*); still other terms express water in the form of *rain*, *dew*, *wave*, and *foam*. It is perfectly conceivable that this variety of ideas, each of which is expressed by a single independent term in English, might be expressed in other languages by derivations from the same term. Another example of the same kind, the words for *snow* in Eskimo, may be given. Here we find one word, aput, expressing *snow on the ground*; another one, qana, *falling snow*; a third one, piqsirpoq, *drifting snow*; and a fourth one, qimuqsuq, *a snowdrift*".

Most of the languages have their own ways to express colors and numbers, but the existence of cross-linguistic universals categories is contested. Kay and Regier (2003) present a recent analysis of color categorization among languages of industrialized and nonindustrialized societies. The results suggest that color categories may not be universal. Everett (2005) described a language called Pirahã, spoken by an indigenous people of Amazonas, Brazil, which lacks such fine categorization, all they count on is the distinction between dark and bright (for colors), few and many (for numerals). Although they lack the semantic representation of these concepts, they appear to understand that there are different concepts. "A total lack of exact quantity language did not prevent the Pirahã from accurately performing a task which relied on the exact numerical equivalence of large sets"(Frank et al., 2008). Even if there are different categorizations in a culture, it might not be reflected on their expression as a language. Also, the existence of different words to acknowledge *snow* in various contexts doesn't imply that Eskimo are more concerned about snow than then the residents of Alaska who speak English. The extent to which language and thought are bound together is fuzzy and unclear.

Consider the simple example of the resemblances that ordinal numbers have to cardinal numbers. In most of the languages this similarity is present in all numerals but the first. English presents different patterns for its first three ordinals and cardinals. See the examples bellow where we have underlined the numerals that don't present resemblances between their ordinal and cardinal expressions:

English: one (first), two (second), three (third), four (fourth), five (fifth), six (sixth), seven (seventh), eight (eighth), nine (ninth)

German: eins (erste), zwei (zweite), drei (dritte), vier (vierte), fünf (fünfte), sechs (sechste), sieben (siebte), acht (achte), neun (neunte)

French: un (premier), deux (deuxième), trois (troisième), quatre (quatrième), cinq (cinquième), six (sixième), sept (septième), huit (huitième), neuf (neuvième)

Hebrew: ehad (rishon), shnayim (sheni), shlosha (shlishi), arba'a (revi'i), hamisha (hamishi), shisha (shishi), shiv'a (shvi'i), shmona (shmini), tish'a (tshi'i)

Estonian: üks (esimene), kaks (teine), kolm (kolmas), neli (neljas), viis (viies), kuus (kuues), seitse (seitsmes), kaheksa (kaheksas), üheksa (üheksas)

There seems to be no explanation for this observation, there is no reason why some languages would behave differently from others when regarding numerals names. It also doesn't seem plausible that some cultures have different perception with respect to ordering things that could create such discrepancy. It might than be that languages are the

way they are just by chance, but it is highly compelling to try to find reason for some common behavior.

We structure our thought using our language, but there are certain situations where words lack and still exist thought, although it might not be represented as a beautiful chain of the language symbols. (Pinker, 2003) discusses several examples of thoughts that do not appear to be represented in the mind in anything like verbal language.

An American author, political activist, and lecturer, named Helen Adams Keller was born in 1880. When she was 19 months old, she contracted an illness which might have been scarlet fever or meningitis. It left her deaf and blind. Despite all difficulties imposed by her impairments she was taught and today she is widely known as the first deaf-blind person to earn a Bachelor of Arts degree. "The story of how Keller's teacher, Anne Sullivan, broke through the isolation imposed by a near complete lack of language, allowing the girl to blossom as she learned to communicate, has become known worldwide through the dramatic depictions of the play and film *The Miracle Worker*". At age 22, Keller published her autobiography, *The Story of My Life* (1903), from which the excerpts bellow are taken from:

> Have you ever been at sea in a dense fog, when it seemed as if a tangible white darkness shut you in, and the great ship, tense and anxious, groped her way toward the shore with plummet and sounding-line, and you waited with beating heart for something to happen? I was like that ship before my education began, only I was without compass or sounding-line, and had no way of knowing how near the harbour was. "Light! give me light!" was the wordless cry of my soul, and the light of love shone on me in that very hour.
>
> (...)
>
> We walked down the path to the well-house, attracted by the fragrance of the honeysuckle with which it was covered. Some one was drawing water and my teacher placed my hand under the spout. As the cool stream gushed over one hand she spelled into the other the word water, first slowly, then rapidly. I stood still, my whole attention fixed upon the motions of her fingers. Suddenly I felt a misty consciousness as of something forgotten – a thrill of returning thought; and somehow the mystery of language was revealed to me. I knew then that "w-a-t-e-r" meant the wonderful cool something that was flowing over my hand. That living word awakened my soul, gave it light, hope, joy, set it free! There were barriers still, it is true, but barriers that could in time be swept away. I left the well-house eager to learn. Everything had a name, and each name gave birth to a new thought. As we returned to the house

> every object which I touched seemed to quiver with life. That was because I saw everything with the strange, new sight that had come to me.

"Language is our legacy. It is the main evolutionary contribution of humans, and perhaps the most interesting trait that has emerged in the past 500 million years" (Nowak et al., 2002). Understanding the origins of our syntactic communication process, how language works and evolves is important to understand our own legacy. Currently there are many efforts from multidisciplinary fields of study, inquiring and seeking to understand and better describe language. The study of formal language theory, learning theory, human psychology and physiology, observations and empirical validations, statistics and mathematical modeling are some of the aspects taken under consideration when studying language as a biological phenomenon of communication.

3

Language Structure

"The poet, the philologist, the philosopher, and their traditional associates have ceased to be the only ones concerned with the structure of human discourse. Increasing numbers of mathematicians and of engineers have joined them, attracted by technological problems, formerly ignored or considered trivial, suddenly become so important that they elide traditional barriers between fields, even those which merely contained intellectual curiosity" (Mandelbrot, 1965). The study of language has become an interdisciplinary field where contributions from different points of views might be important to draw deeper insights on the understanding of this communication phenomena. It is a difficult task to formalize a theory to explain how order is able to emerge from the unpredictability characteristic of the human discourse.

To understand the communication process through language, it is necessary to build a model to explain how languages work. This model should be able to explain how we use language to code information into utterances and how we use our language knowledge to interpret an incoming spoken message and retrieve information. Those aspects are under concern of phonology. A simple explanation of what regards phonology is to say that it stands for the study of sound structure in language. The object of study of phonology is "an abstract cognitive system dealing with rules in a mental grammar: principles of subconscious *thought* as they relate to language sound. (...) the *sounds* which phonology is concerned with are symbolic sounds – they are cognitive abstractions, which represent

but are not the same as physical sounds"(Odden, 2005). In the study of phonology we are concerned about what are the abstract representations of language and how are the rules in the coding process, from an abstract entity into a physical realization manifested as acoustic waveform, which carries intrinsic information of the speech utterance, i.e. information of the message, code and medium.

Not every type of sound may be produced by our phonatory apparatus. Even posed that restriction, the repertoire of speech sounds is enormous (919 different speech sounds were found in UCLA Phonological Segment Inventory Database, UPSID). The sounds of languages are also not equal across the multitude of languages in the World. In German the word for 'beautiful' is 'schön' (pronounced as [ʃøːn]). The vowel [ø] does not exist in English, Portuguese and other languages, but it does in French (the word 'jeûne', pronounced as [ʒøn], meaning noun 'fast') and Norwegian (the word 'øl' means 'beer' and is pronounced as [øl]), for example. Not only the sounds differ from language to language, but also the way they might be combined. Certain combinations of sounds are allowed and others are impossible in a language. Some combinations are allowed in certain positions, but not in others. In English, the consonantal cluster [ts] may not happen word initially; there is no English word beginning with [ts], but it may be found in word final position, such as 'cats' ([kæts]) and 'splits' ([splits]). In German, the consonantal cluster is allowed in word final position, as the word 'Konferenz' ([kɔnfe'rɛnts]), meaning 'conference'; and it is also allowed word initially, the word for 'time' is 'Zeit' ([tsaɪt]).

The knowledge of those rules is somehow internalized in a native speaker abilities so that (s)he may judge weather or not an unknown word fits the language structure, meaning that this unknown word could be a possible or impossible word in that language. It is usual to find people playing of making new words.

> I have also invented some new words. 'Confuzzled', which is being confused and puzzled at the same time, 'snirt', which is a cross between snow and dirt, and 'smushables', which are squashed groceries you find at the bottom of the bag. I have sent a letter to the Oxford Dictionary people asking them to include my words but I have not heard back. – from the movie 'Mary and Max' (2009)

It would not be a surprise to find those invented words in a dictionary someday, since every word is an invented linguistic sign. Some examples that have been recently incorporated to English are 'robotics' (meaning the technology of design, construction, and operation of robots) invented by Isaac Asimov[1] in his 1941 science fiction short-story 'Liar!'; and 'warp

[1]Isaak Yudovich Ozimov was born in Petrovichi, Byelorussian SSR, between October 4, 1919 and January 2, 1920 (this date is uncertain due to the lack of records, of the Jewish and Julian calendars). His family immigrated to the United States when he was three years old. He lived in the Brooklyn, New

speed' (meaning the highest possible speed), term first used in the 1960s by the series *Star Trek* and incorporated in the dictionaries in 1979. The term 'snirt', although not yet incorporated to many dictionaries, is already reported by more dynamic dictionaries like the *Wikitionary* and is commonly used by people. In German, that is an analytical language, this kind of dirty snow is named 'Schneematsch' (comes from 'Schnee' (snow) + 'Matsch' (mud)).

"In particular, the frequency with which individual words or sequences of words are used and the frequency with which certain patterns recur in a language affect the nature of mental representation and in some cases the actual phonetic shape of words"(Bybee, 2003). The structuralism provided a way to analyze the speech continuum into a sequence of units, and these units into features; establishing hierarchical relations between them and organizing the speech knowledge into different levels of a grammar built of phonology, morphology, syntax, and semantics. The way language is used as a social-interactional tool and the frequency of occurrence of certain patterns are determinant factors to explain the language phenomenon. "It is certainly possible that the way language is used affects the way it is represented cognitively, and thus the way it is structured"(Bybee, 2003).

"The proposal that frequency of use affects representation suggests a very different view of lexical storage and its interaction with other aspects of the grammar or phonology than that assumed in most current theories. Structuralist and generative theories assume that the lexicon is a static list, and that neither the rules nor the lexical forms of a language are changed at all by instances of use"(Bybee, 2003). It is important to have a language model capable of explaining some language usage facts as, for example, the fact that the rate and extent of a phonological change is directly affected by the frequency of the involved items in the lexicon. The way phonological rules and phonological representations are stated should consider those aspects of languages. A good conceptualization of phonology shall not forget that, as part of the procedure for producing and understanding language, the phonological properties of a language must be highly associated with the vocal tract and its usage. Bybee (2003) proposes that language is governed by cognitive and psychological processes and principles which are not language specific, but the same that govern other aspects of human cognitive and social behavior.

We believe language might be modelled as a self-organized complex system which is made up of many interacting parts. Each individual part, called 'agent', interact in a simple way with its neighbours, what might lead to large-scale behaviours that might not be predicted from the knowledge only of the behaviour of the individual parts. What we observe as phonological rules and patterns of language might be understood as these col-

York. He studied in the Seth Low Junior College for two years and then in the Columbia University, where he graduated in 1939. After he made a Ph.D. in biochemistry in the same institution. In his life, Asimov wrote and edited more than 500 books and is widely known by his science-fiction writings.

lective behaviours that emerge from this simple process of interacting agents in a complex system. The interactions between agents might be taken in different forms. The famous predator-prey approach is used by Wang et al. (2004) to model lexical diffusion. Such computational studies of language emergence provide a valuable way to study language evolution. This field of study started with Hurford (1989) simulation model on lexical emergence and acquisition. Hurford considered "three conceivable strategies for acquiring the basis of communicative behaviour, here labelled the Imitator, Calculator, and Saussurean strategies" (Hurford, 1989). The first strategy consists of imitating others agents when they refer to certain objects. The second approach consist of reinforcing the usage of a certain utterance when the neighbour agents respond positively. According to Hurford, the better approach is the last one, the Saussurean strategy, in which the agents copy the patterns produced by nearby individuals, but make their perception consistent with their own production. "Many models of language evolution have adopted the agent-base simulation paradigm" (Wang et al., 2004). All of them need real world quantitative data to use on the models and to validate the outcomes. On the following sections we intend to quantitatively analyze some natural language data.

4

The Phonemic Model of Language

> En matière de langue on s'est toujours contenté d'opérer sur des unités mal définies (In language's matter it has always been sufficient to operate on ill-defined units).
>
> Ferdinand de Saussure (1916)

The sound structure of languages is under the focus of phonology. Under the phonological point of view, it is not restricted to the sounds at the physical level of speech realization, but also extended to the symbolic 'sounds', which are cognitive abstractions that provide a mean of representation for a language.

The 'sound' as a physical phenomenon is a complex pattern of disturbance that travels across all forms of matter: gases, liquids, solids, and plasma; in this context, called the medium. In speech communication, the sound source, the emitter, through its oral tract, produces a disturbance that travels in the free air medium. This disturbance suffers scattering, attenuation and other sorts of distortion as it travels across the medium. The striking force of the disturbance reaches the receptor's ear, what causes, after a series of transformations in the outer, middle and inner ear, neural signals that are sent to the brain and perceived as speech sounds, eliciting meaning and creating communication.

In order to establish a phonemic model of language it is necessary to define which

are those cognitive symbols and the rules under which they interact to create a cognitive representation of speech. The Swiss linguist Ferdinand de Saussure is responsible for shifting the way linguistics was done and established a break point with his posthumous publication *Course in General Linguistics* in 1916. Structural linguistics was the new approach to linguists and the creation of phoneme was the basis for all the new born linguistics. To collect a corpus of utterances and to produce an attempt to classify all the elements of the corpus at their different linguistic levels was the new paradigm that brought linguistics from diachronic to synchronic analysis.

As pointed out by Capek (1983): "Saussure did not discover – or invent – the phoneme. He was one of a number of scholars working on comparable ideas. His statements are important not so much for the accuracy with which they report facts or for their strength and consistency as for the explanatory power of the inferential network they set up. Suffice it to say that after Saussure the clumsy terminology used by scholars who preceded him – that of speech sounds (*Sprachlaute*) – came to be replaced by the much more intuitively satisfactory concept of the phoneme."

The idea of phoneme comes from the idea of minimal differences that creates meaning. It is regarded as the smallest segment unit in speech that makes meaningful contrast between two utterances. Each language has its own phoneme inventory, for example, in English /t/ and /d/ are distinct phonemes because a simple change from /to/ to /do/ creates a meaning difference. The languages of the world have different sets of phonemes, or, more appropriately, we may say that they have different categorizations. In English no distinction is made between aspirated and unaspirated sounds. In Hindi such distinction exists: /tal/ (beat) contrasts with /t^hal/ (plate) (Ladefoged and Maddieson, 1996). Implicit in the idea of phoneme is that language may be segmented into a sequence of consecutive symbols in time; those segments may be differentiated one from the other; and different symbols create the possibility of different meaning, according to the context in which they are inserted. Benoît Mandelbrot showed in his paper (Apostel et al., 1957) that speech, in order to be comprehensible in the most diverse situations and under heavy corruption, must necessarily, at a certain level, be understood as a discrete process, because only in this manner it would make speech comprehension possible in such degenerated situations. Another advantage of the discrete aspect of language is that discreteness at the phonetic level guarantees the discreteness at all other levels, and that is the base of all linguistic knowledge nowadays.

"Man could not perceive speech well if each phoneme were cued by a unit sound"(Liberman et al., 1967). The acoustic cues in an utterance carry information in parallel about successive abstract units, the phonemes. It builds a complex relation between cues and phonemes, and it makes speech perceiving rate slower but also more robust. Liber-

man et al. (1967) pointed some reasons why a speech code could not be alphabetic, a one-to-one correspondence between codes and phonemes. If man can follow a speech at a rate of 400 words per minute, and if each word has an average of 4 to 5 phonemes, that would lead to 30 phonemes per second, what would overrun the human temporal resolving power of sound stimuli (Miller and Taylor, 1948). It evidences the necessity to have a surjective mapping[1] from acoustic cues into phonological abstract elements. Other acoustic alphabetical codes used for communication have shown that it is hard to achieve a communication rate near the speech rate. The simple example of Morse code shows how ineffective this type of communication is, where the highest rate ever achieved by a skilled operator was 75.2 words per minute (Pierpont, 2002). Other codes, used in reading devices for the blind (the Optophone, built by Fournier d'Albe in 1913; the Optacon, by John Linvill in 1962; the Stereotoner in 1973 by the Mauch Laboratories; among other examples) were tested, and none showed a performance near the speech communication rate. Another drawback posed by the alphabetic approach is the difficulty of identification of numerous and distinct acoustic stimuli. The works of Miller (1956); Pollack (1952) among others suggest that this number is considerably less than 31, which is the average number of phonemes in a language (the language with more phonemes is !Xu, spoken in southern Africa, with 141 phonemes; and the languages with fewer phonemes have only 10, the Pirahã, spoken by indigenous people in Brazil, and the Rotokas, spoken by a few people in Bougainville, an island to the east of New Guinea (Maddieson, 1984)).

The chief problem in determining the form and behavior of phonemes in a certain language system is to achieve a method of quantitative comparison between two or more phonemes. We may describe speech sounds by their place and manner of articulation, we may extract some of their acoustic characteristics and we may determine whether two sounds build a phonemic contrast in a language, but it is still hard to establish a significant quantitative dissimilarity measure of two phonemes and then discover a unique scale under which all phonemes might be measured.

Any speech sound production may be described as a sequence of articulatory gestures. Imagine that we take an X-ray moving picture of a person speaking and build a slow motion picture of the activity of the speech organs during the utterance production. We might then describe in details the articulatory gestures. Suppose for example we are analyzing the utterance of a [t]. We might see the tip of the tongue rising to the top of the back of the upper teeth, forming an occlusion and releasing it. If we analyze the utterance of a [d] instead, the description of the gestures would be quite similar, but added by a contraction, vibration and relaxation of the vocal cords to produce voicing.

As pointed out by Zipf (1949), "the point of concern at present, however, is not to

[1]Surjective mapping is a map from one set onto another set, so that its range is the entire second set.

devise a system of symbols whereby the sequence of sub-gestures constituting a speech-sound can be noted, but rather to remark that speech-sounds and phonemes may be viewed as constellations, or configurations, of articulatory sub-gestures, arranged partly or completely in sequential order, some sequences running concurrently with others (e.g. the occlusion is concurrent with the voicing of *d*). Although there is not a single speech-sound, or variant form, of a phoneme in any language which cannot be conceived of as a constellation or configuration of the type envisaged above, yet a complete and accurate description of even a single speech-sound in terms of sequences is practically impossible. However, by conceiving phonemes as constellations of this order, we have found a very useful method of comparing a number of specific phoneme pairs which have essential sequences in common."

Zipf (1949) proposes that the frequency a phoneme in a language occurs is inversely proportional to its complexity. He analyzed aspirated and unaspirated stops in four languages: Peipingese Chinese, Danish, Cantonese Chinese and Burmese. The data analyzed corroborated his hypothesis. He also analyzed the occurrence of voiced and voiceless stops in twelve languages (Czechish, Dutch, French, Italian, English, Hungarian, Bulgarian, Russian, Spanish, Greek, Latin and Sanskrit) and concluded that the voiceless stop is more frequent then its voiced counterpart. When he compared the occurrence (in German and Sanskrit) of long vowels against short vowels and diphthongs against each of its parts alone, he once again found that the simple one (short single vowel) has a greater frequency of occurrence compared to the complex counterpart (long vowel or diphthong). Comparing also the occurrence of [m] and [n] across 21 languages, he observed that [n] was more frequent. "It might be taken as some evidence that *n* is the simplest of the two phonemes because of the observation of comparative philology which indicates that quite often, when *m* disappears in any of its usages in a given language, it becomes (i.e. 'weakens' to) *n* before disappearing"(Zipf, 1949). A more complete and recent analysis (Maddieson, 1984) shows that the observations of Zipf might be wrong. In a database of 425 languages, the phone [m] is present in 94.2% of the languages, against only 44.8% for the phone [n] and 16.9% for those languages that have both [m] and [n]. The explanation could then be the other way around: [m] is simpler than [n], and it gets more complex before disappearing (it is preferable to keep in the speech repertoire only simple and contrastive symbols). The only way to corroborate this kind of analysis would be by using the information on the usage frequency of these phones, which is unfortunately unavailable. Haplology[2] and the Grassmann's law[3] are cited by Zipf (1949) to exemplify

[2]Haplology is the elimination of a syllable when two consecutive syllables are identical or similar.

[3]The Grassmann's law is a dissimilatory phonological process observed in Ancient Greek and Sanskrit. According to this law, if an aspirated consonant is followed, in the next syllable, by another aspirated consonant, the first one loses the aspiration, reducing the complexity.

the relation between the change in the pronunciation complexity and the change in its frequency of occurrence.

In the history of any language, the phonemic system undergoes constant changes that may affect the complexity and occurrence frequency of phonemes. Comparing two different languages with a common ancestral we might notice these changes and their consequences on frequency and usage. As it has been quite generally observed, the emergence of a phonetic change in a certain language's phonemic system is unpredictable. Any phoneme may find itself suddenly unstable and undergoes a change, whether in all occurrences of this phoneme or, for example, only in certain situations restricted to the relative position it presents. These phonemes that suffered change might stay stable or suffer a subsequent change. According to Zipf (1949), "though the emergence of phonetic change is apparently capricious, actual phonetic changes seem to fall into four main and orderly types: (1) the so-called *'spontaneous' changes*; (2) the *accentual changes*; (3) the *assimilatory changes*; and (4) the *dissimilatory changes*.

A comparative analysis on the articulation of many given phonemes and phoneme sequences shows that some phonemes are facilitated due to the restrictions on the physiology of the mouth and a contiguous speech tendency. As an example, we clearly understand why it is easier to pronounce a [d] after [n] then after [m]. When we pronounce an [n], the tongue touches the hard palate just behind the upper teeth and that is also the articulation point for [d]. The pronunciation of [m], on the other hand, requires a bilabial closure, what is not part of the articulation for [d]. This facilitation process is clearly observed when we analyze the statistics of diphone occurrences in a language. In English, for example, the occurrence probability of a [d] given that a [n] has occurred adjacently is 34.9%, much greater than the 1.8% probability of occurring the same [d] given that an [m] has just occurred. When comparing a pair like [ts] and its reversal [st], the observed probabilities of occurrence in English are 5.1% and 33.3%, respectively. The only difference between these pairs is in the movement of the tongue tip.

Although we don't have the means to acquire the same sort of information for all languages in the world, we might use the UPSID database (Maddieson, 1984) to verify what is the probability of occurrence of a phone given that the other is present in that language. Concerning the three phones in the previous analysis [m,n,d], we get the following result: 94.2% of the languages that have [d] in its repertoire also have [m] in it; 89.2% of the languages that have [d] in its repertoire also have [n] in it; 26.6% of the languages that have [m] in its repertoire also have [d] in it; 53.0% of the languages that have [n] in its repertoire also have [d] in it; 47.0% of the languages that have [m] in its repertoire also have [n] in it; and 99.0% of the languages that have [n] in its repertoire also have [m] in it.

As another example, consider the occurrence of [t] and [d] between vowels. It is easier to keep the vocal folds vibrating during the whole interval, producing a [d], than stopping its vibration for a very short time and starting it again, producing then a [t]. Zipf (1949) gives as an example the pair [p]-[b]. The Latin word for 'river bank' is *ripa*, which became *riba* in Old Provençal, due to this facilitation process. Zipf (1949) argues that "when an intervocalic *p* becomes voiced in a given language, it is an indication rather of the instability of *p* in that given language that of a universal instability of intervocalic *p*; for example, we have for centuries been pronouncing intervocalic *p* in English *rapid*, *tepid*, *paper* without the shifting of *p* to *b*". Comparing the statistics of occurrence of VCV-like triphones in English, the percentage of VtV is 10.13% against 5.59% of VdV; VkV 8.59% against 2.18% for VgV; VfV appears 8.36% and VvV 3.88%; VbV 5.02% almost like its voiceless counterpart VpV with 4.70%. In all the cases presented, except the bilabial one (which shows a very similar frequency of occurrence), the voiceless consonant exhibits a distinguished frequency of occurrence. This result shows a greater stability of the voiceless consonants in English, although it is known that in some dialects of American English there is a tendency towards voicing the intervocalic plosive consonant [t]. "In these dialects the differences between *latter* and *ladder*, *kitty* and *kiddy*, *metal* and *medal* are practically subliminal; the couplets are distinguished more by the usage of the words than by perceptible differences in the phonemes"(Zipf, 1949).

Historical linguistic analysis shows that the Latin[4] word *ripa* became *riba*, and from that came the word *rive* in French. The Latin *faba* became the French *fève*, "the intervocalic *p* of Latin began to approximate the norm of *b* to such an extent that in intervocalic positions the two behaved indistinguishably, and subsequently, in losing their explosiveness in this position, both began to approximate the norm of *v*"(Zipf, 1949). This process of shift suffered by some phonemes is such a slow process that has to be apprised over a considerable extent of time. The statistical analysis of relative frequencies is important to determine the dynamics and evolution of a language. The balance between phonemes and their metamorphosis through time are keys to understand how linguistics categories are built and shaped. It is important to study the processes of assimilation (merge) and differentiation (split) in a language. As pointed by Zipf (1949), "every assimilation points to a weakening or instability of the assimilated sound, and this weakening or instability is caused primarily by the excessive relative frequency of the assimilated sound".

One example of differentiation process is found in Germanic languages where the back vowels [u] and [o] were originally in an allophonic relation to [y] and [ø], respectively,

[4]Nowadays there are two conventions on the pronunciation of Latin: the *Church Latin* developed out of Medieval traditions and used as the standard pronunciation in the Roman Catholic Church, as a virtually living language; and the *Classical Latin*, spoken by the educated Romans of the late Republic and Empire periods, which pronunciation was reconstructed by historical linguistics of the 19th Century

when they are placed before a following vowel /i/. With time, the syllables containing this /i/ were lost, and a phonemic split made the phones /y/ and /ø/ distinct phonemes. Two front rounded phonemes were added to the vowel system repertoire. Analyzing the vowel system in various languages, we observe that most of them have front unrounded vowels or diphthongs (449 languages of all 451 languages in the UPSID database) and only very few languages have front rounded vowels or diphthongs (46 languages in the UPSID database). This means that only 2 languages have a vowel system built exclusively on front rounded vowels. In the same way, there are 44 languages using a mixed vowel system with front rounded and front unrounded vowels; and 405 languages using an exclusive front unrounded vowels system. An explanation for the reason of this finding is that most languages choose vowels that are maximally distant from one another. Front vowels have a higher second formant (F2) then back vowels; unrounded vowels also have a higher F2 than rounded vowels. This means that unrounded front vowels and rounded back vowels have maximally different second formant, which enhances the differentiation between them. Every dissimilation points to a strengthening of the existing splitting sounds, their excessive relative frequency and their semantic difference build a barrier to create distinction where, before, was a categorical amalgam.

We might wonder that there could exist thresholds of tolerance comprising the relative frequency of phonemes in order to justify or deny the applicability of processes of assimilation or differentiation. Those thresholds are limits to the relative frequency of a phoneme, bellow which a phoneme tends to weaken and above which it tends to strengthen, creating merges and splits. At a first analysis of what might be these thresholds, we have to keep in mind an obvious fact: a sufficient variability is important to create distinct symbols, so that the permutation of them can express the information exchanged during communication. Among the 451 languages in the UPSID database, the minimum number of segments used by a language is 11, in only two languages (Pirahã, spoken in Brazil by around 300 speakers; and Rotokas, spoken in Papua New Guinea by approximately 4,300 speakers). On the other hand, it is also not efficient to have a code with too many symbols, because the probability of false detection would increase so much and make communication unproductive. The language with the highest number of segments found in the UPSID database has 141 segments (the !Xu language, also called !Kung, is spoken by fifteen thousand speakers in Namibia and Angola). On average, the languages are built on 31 segments. Figure 4.1 shows the histogram of languages regarding the number of segments used in each of them.

The analysis using the CMU Pronouncing Dictionary of English shows that this language has 39 phonemes, not counting variations due to lexical stress. The frequency of occurrence of these phones is presented in Figure 4.2. It shows that the most frequent

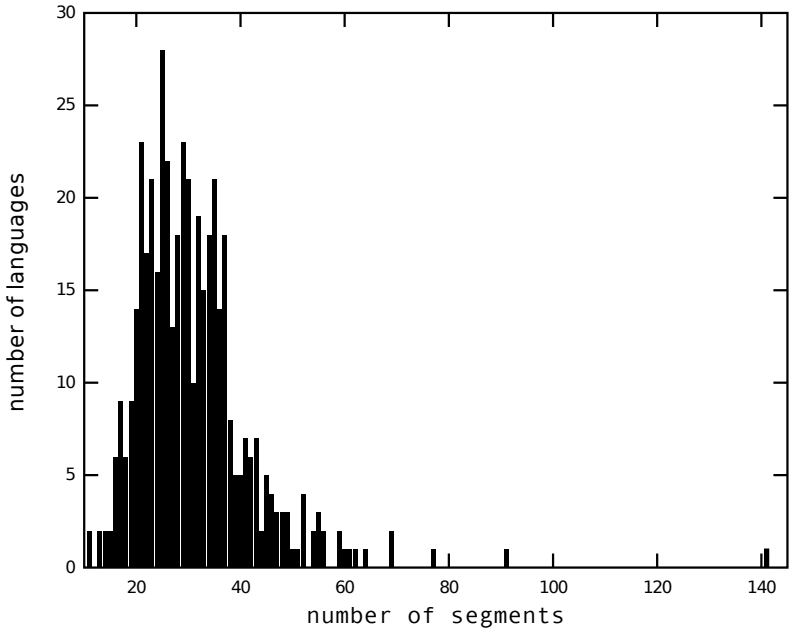


Figure 4.1: Number of segments in various languages of the world, using the UPSID databese.

symbol is [ə], accounting 10.12% of symbols occurrence in that language. Following in order of frequency of occurrence comes [t] with 6.99% and [n] with 6.92%. The most infrequent phoneme in English is the diphthong [ɔɪ] which occurs with a relative frequency of only 0.1%. Preceding it is the phoneme [ʊ] with 0.42% of relative frequency (see chapter 6 for more information).

If we had these relative frequencies, during one language evolution, like the previous ones derived for English, they could be used as a first approximation to determine the thresholds for assimilation and dissimilation processes. The phonemes cannot have, all of them, the same percentage-threshold, for the weakening of a phoneme, causing an assimilation process, could cause the target phoneme to an excessive relative frequency, and it would require this target phoneme to be capable of sustaining a higher frequency than the vanishing one. It is quite reasonable to consider that phonemes have distinct thresholds, what might be explained by their different acoustic and usage properties. It would be necessary to apprise and compare the features of phonemes, and also the way they are used and connected to build utterances.

According to Zipf (1949), when a phoneme become so rare, with a relative frequency abnormally low, "the phoneme then would become a distinctive and very characteristic part of every word in which it occurred". In such situations, an accidental epenthesis might appear, like the strengthening of [t] into [ts] observed in the Old-High-German sound-shift in which a Germanic [t] shifted to a [ts] in the majority of positions, and it went even further in some cases to [ss]. Some examples in German, compared to the

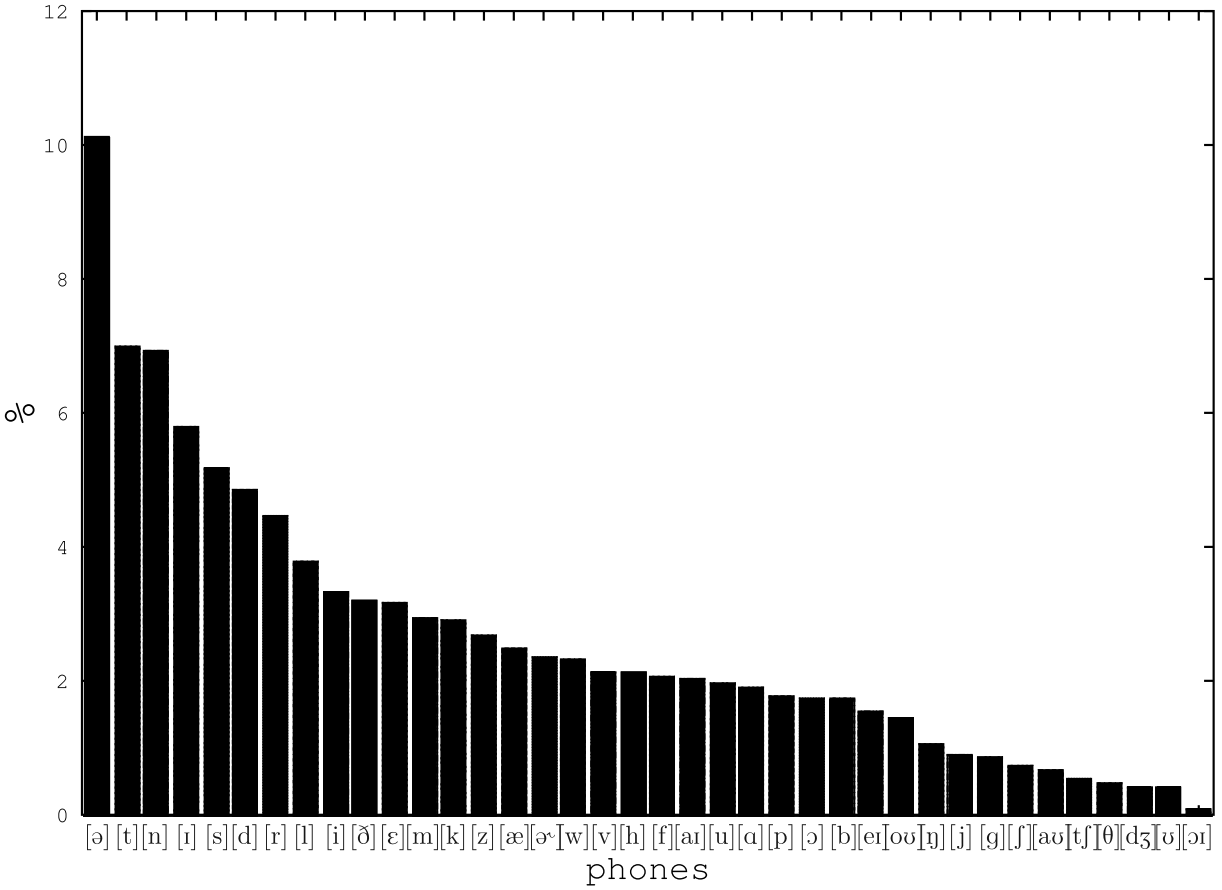


Figure 4.2: Frequency of occurrence of English phones.

English counterpart, which preserved the [t], are: *zwei*, *two*; *zehn*, *ten*; *zug*, *tug*; *zahn*, *tooth*; *zeit*, *time*.

Observing 12 languages, Zipf (1949) concludes that, with a few exceptions (Spanish *d* and *t*, and Hungarian *b* and *p*), the voiceless stops strongly outnumber their voiced counterparts and the relative frequencies of occurrence are amazingly similar. Considering that the voicing assimilation is a normal process found in many languages, it seems a quite astonishing result. It seems like other factors prevent this assimilation process from occurring. We could then consider the hypothesis that voiced stops have a lower threshold, and the assimilation process would force them to cross this threshold. In this situation, the assimilation would not move forward, and the voiceless stops are in a great number preserved.

The temporal nature of speech is evident. During an utterance, a speech context is built, and such a context is capable of influencing how speech is perceived. The perception of an initial sound of [k] or [g] is influenced by the following sounds, [ɪs] or [ɪft], for example, creating confusion and leading to false identifications (in the case of [kɪft] or [gɪs]) due to the existence of the words 'gift' and 'kiss'. The perception of what is currently being

uttered both influences and is influenced by the perception of what comes later and what came previously.

Another aspect, as we observe speech phenomena through time, is that, although we fight to achieve a segmental model of successive speech units, it has not been shown possible to split the speech continuum into a sequence of discrete elements, since the speech cues frequently overlap in time. The schematic problem of the overlap is shown in Figure 4.3. The problem of overlap is less severe but still exists at word level. In normal speech, mainly in rapid speech, words run into each other. This phenomenon is easy to perceive when we are listening to a foreign language and we can't tell when a word ends and another starts. It is not unusual to have speech errors due to wrong segmentation of words, what might be influenced by the context (Bond and Garnes, 1980; Cole and Jakimik, 1980).

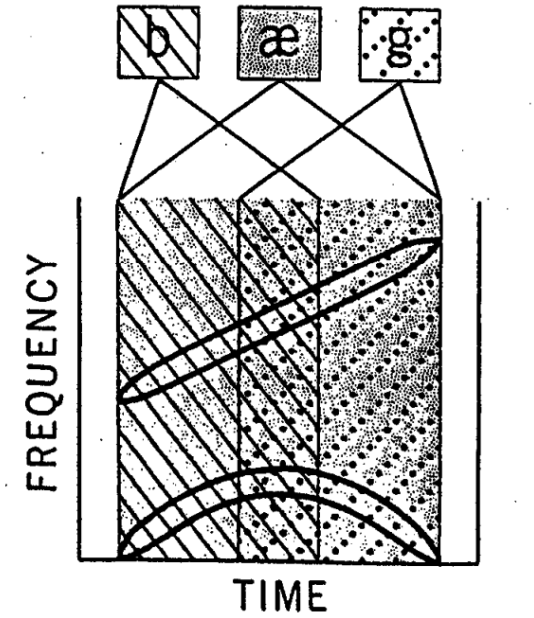


Figure 4.3: A schematic spectrogram for the word 'bag' presenting the overlapping of acoustic cues. (Liberman, 1970)

There have been many ideas of how speech perception works and what are the basic units of it. It is still not clear how we segment speech as we perceive or whether we segment it at all. Various researchers advocate in favor of different approaches: Klatt (1979) arguments in favor of diphones; Pisoni (1982) in favor of phonemes, what seems to be the most accepted view in the literacy; Fujimura and Lovins (1978) argues in favor of demisyllable; Wickelgren (1969) is in favor of context-sensitive allophones; and Studdert-Kennedy (1976) is in favour of using syllables as basic units. It seems reasonable to take the advantages of each approach and overcome each of their drawbacks, to take the problem under a multi-resolution point of view. It would, at a first glance, increase the

complexity of the model.

The cues not only overlap one into another, but they also are different according to the various context they may appear. As presented by Liberman et al. (1967), the second-formant transition is responsible for determining what consonant the listener perceives. Figure 4.4 presents some results for different transitional patterns for the second-formant.

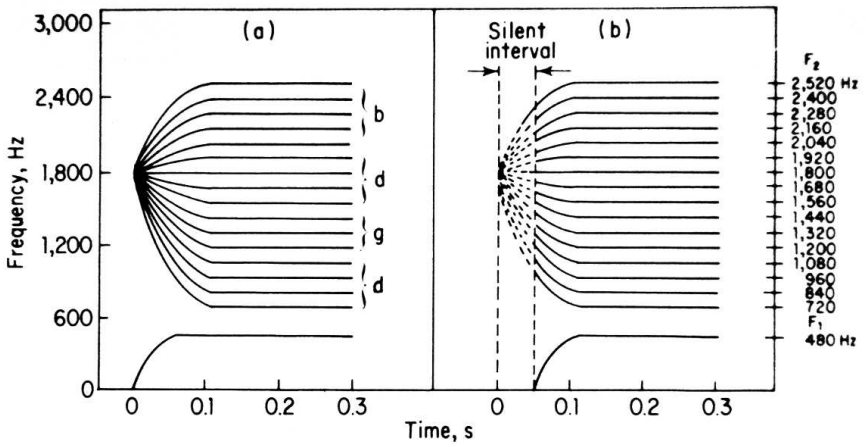


Figure 4.4: The second-formant starts at, or points to, the /d/ locus. On (a), the syllables were perceived as /b/, /d/ or /g/, depending on the final frequency level of the formant. On (b), all syllables were perceived as beginning with /d/. Figure taken from Dalattre et al. (1955)

5

Pronouncing Dictionary

> People are under the impression that dictionaries legislate language. What a dictionary does is keep track of usages over time.
>
> Steven Pinker

Pronunciation dictionaries are lists of words or phrases with their respective pronunciation transcribed into a phonetic alphabet. The pronunciation of words may vary much in spontaneous speech, and for that reason many dictionaries include some of the possible variations found in spoken interactions. Pronunciation dictionaries usually reflect one particular spoken accent, usually chosen as the most neutral among the various accents in a language. They are constructed by hand or by a rule-based system. Pronunciation dictionaries are most used in speech recognition system and synthesizers, and the usage of an appropriate one may improve significantly the system performance (Lamel and Adda, 1996). Research efforts have been aiming to build pronouncing dictionaries that are automatically trained with real speech data and preliminary experiments have shown the achievement of good results, eliciting higher recognition rate systems (Fukada and Sagisaka, 1997).

In order to acquire statistical information on speech pronunciation using a text database, we may use a pronouncing dictionary to transcribe words into a sequence of phones. A

few tools are available nowadays to be used in this purpose:

Moby Pronunciator II contains 177,267 words with corresponding pronunciations fully International Phonetic Alphabet coded. Stress or emphasis is also marked in the data. It was createb by William Grady Ward and in 2007 has been placed into the public domain.

TIMIT corpus of read speech contains a total of 630 sentences spoken by 10 different speakers from 8 major dialect regions of the United States. The total number of words in the corpus is only 659 words. The corpus has phonemically and lexically transcribed speech. It was created as a joint effort from the Massachusetts Institute of Technology, Stanford Research Institute, and Texas Instruments.

ICSI Switchboard is a corpus of several informal speech conversations, containing over 3 million words, recorded over the telephone. It includes a pronouncing lexicon with 71,100 entries using a modified Prolex phonetic alphabet.

CMUdict is a public domain dictionary created by Carnegie Mellon University. It contains 133,746 entries of English words mapping between its written form and their North American pronunciations.

These corpora cited above have been designed to provide data for the creation of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems and speech synthesizers.

The present work aims into acquiring statistical knowledge of the patterns found in the acoustic-phonetic behavior of the English language. To reach this sort of information from textual data, we chose to use the Carnegie Mellon University Pronouncing Dictionary (CMUdict) since it is available in public domain, is widely used and has many entries.

5.1 The Carnegie Mellon University Pronouncing Dictionary

The Carnegie Mellon University (CMU) Pronouncing Dictionary is a machine-readable pronunciation dictionary for North American English created as public domain resource. It defines a mapping from English words to their North American phonetic transcriptions. Those transcriptions are coded by the ARPAbet[1] (Shoup, 1980), a phonetic transcription

[1] "Arpabet is a phonetic transcription code developed by the Advanced Research Projects Agency (ARPA) as a part of their Speech Understanding Project (1971–1976). It represents each phoneme of General American English with a distinct sequence of ASCII characters. Arpabet has been used in several speech synthesizers, like SAM for the Commodore 64, Say for the Amiga and TextAssist for the PC. It is also used in the CMU Pronouncing Dictionary"(Wikipedia).

code developed by Advanced Research Projects Agency (ARPA). "In November of 1971, the Information Processing Technology Office of the Advanced Research Projects Agency of the Department of Defense (ARPA) initiated a five-year research and development program with the objective of obtaining a breakthrough in speech understanding capability that could then be used toward the development of practical man-machine communication systems. (...) The objectives were to develop several speech understanding systems that accept continuous speech from many cooperative speakers of a General American dialect" (Klatt, 1977). It uses a set of 39 phones that represents the speech inventory of the General American English[2] It doesn't make any kind of surface reduction like flapping or reduced vowels, since "predicting reduction requires knowledge of things outside the lexicon (the prosodic context, rate of speech, etc.)" (Jurafsky and Martin, 2009). Instead, the vowels are marked by a number indicating the stress associated with them.

Tables 5.1 and 5.2 present the equivalence between IPA code and ARPAbet code. The CMU Pronouncing Dictionary was chosen to be used since it is the pronouncing dictionary with the most number of words (contains over 133,000 entries) and it is open and free to use. The CMU Pronouncing Dictionary is also commonly used in many speech processing applications such as the Festival Speech Synthesis System and the CMU Sphinx speech recognition system.

The format used by the CMU Pronouncing Dictionary has mappings from words to their pronunciations using the phone set given by a modified ARPAbet system (the difference is the stress marks used on vowels). The current phone set contains 39 phones (not counting variations due to lexical stress) and the vowels may be marked by their lexical stress using the scale: 0 (no stress), 1 (primary stress) and 2 (secondary stress). When alternate pronunciations exist for a given word, they are marked by an index within parentheses. The first version of the dictionary was release on 16th of September 1993. The version used was 0.7a, released on 19th February 2008.

The analysis proposed consider words as isolated structures and for that reason poslexical phonological rules are not taken in account. Various phonological changes that happen in a continuous speech, such as flapping, vowel reduction, and various coarticulation effects that happen as a postlexical phonological change are not considered, since we are not evaluating the effects of neighbor words on a continuous speech.

[2]General American English, also known as Standard American English, is the standard accent used by most of the American films, TV series, news, advertisements and radio broadcasts. The area of eastern Nebraska, southern and central Iowa, and western Illinois are considered to be the places where local accent is most similar to General American (Labov et al., 2006).

Table 5.1: Arpabet Symbols and their IPA equivalents : Vowels

		IPA Symbol	ARPAbet	Example	Transcription
Vowels	Front	i	IY	beat	B IY1 T
		ɪ	IH	bit	B IH1 T
		ɛ	EH	bet	B EH1 T
		æ	AE	fast	F AE1 S T
	Back	ɑ	AA	father	F AA1 DH ER0
		ɔ	AO	frost	F R AO1 S T
		ʊ	UH	book	B UH1 K
		u	UW	boot	B UW1 T
	Mid	ə	AX	discus	D IH1 S K AX0 S
		ʌ	AH	but	B AH1 T
	Diphthongs	eɪ	EY	bait	B EY1 T
		aɪ	AY	my	M AY1
		aʊ	AW	how	HH AW1
		ɔɪ	OY	boy	B OY1
		oʊ	OW	show	SH OW1
	R-colored	ɝ	ER	her	HH ER0
		ɚ	AXR	father	F AA1 DH ER
		ɛr	EH R	air	EH1 R
		ʊr	UH R	cure	K Y UH1 R
		ɔr	AO R	more	M AO1 R
		ɑr	AA R	large	L AA1 R JH
		ɪr	IH R or IY R	ear	IY1 R
		aʊr	AW R	flower	F L AW1 R

Table 5.2: Arpabet Symbols and their IPA equivalents : Consonants

		IPA Symbol	ARPAbet	Example	Transcription
Stops	Voiced	b	B	bat	B AE1 T
		d	D	deep	D IY1 P
		g	G	go	G OW1
	Unoiced	p	P	pea	P IY1
		t	T	tea	T IY1
		k	K	kick	K IH1 K
Fricatives	Voiced	v	V	very	V EH1 R IY0
		ð	DH	that	DH AE1 T
		z	Z	zebra	Z IY1 B R AH0
		ʒ	ZH	measure	M EH1 ZH ER0
	Unoiced	f	F	five	F AY1 V
		θ	TH	thing	TH IH1 NG
		s	S	say	S EY1
		ʃ	SH	show	SH OW1
Affricates		tʃ	CH	church	CH ER1 CH
		dʒ	JH	just	JH AH1 S T
Nasals		m	M	mom	M AA1 M
		n	N	noon	N UW1 N
		ŋ	NX	sing	S IH1 NG
Liquids		l	L	late	L EY1 T
		r	R	run	R AH1 N
		ɾ	DX	wetter	W EH1 T ER0
Others		h	HH	house	HH AW1 S
		ʔ	Q	glottal stop	
Semivowels		j	Y	yes	Y EH1 S
		w	W	way	W EY1
		ʍ	WH	when	WH EH1 N

6

Language Statistics

> In trying to give an account of the statistical properties of language, one is faced with the problem of having to find the common thread which would show the many and multifarious forms of language statistics - embodied in scattered papers written by linguists, philosophers, mathematicians, engineers, each using his own professional idiom - as belonging to one great whole: quantitative linguistics.
>
> Gustav Herdan (1966)

If we agree to see language as a purpose driven stochastic event, it is important to view its statistics and observe how it is structured. We might stipulate that language is a usage driven process, the way it is used is shaped by its structures, and the way it is structured is dictated by its usage. We see language then as a dynamic process in a constant feedback loop.

To understand the statistical properties and build models of languages is a central task in understanding how language works and create natural language processing sys-

tems. Traditionally, language has been studied and modeled manually, describing each observation and building a language grammar from them. With the recent availability of a massive amount of data, statistically trained models are an attractive alternative. Probabilistic models of language are also fundamental in speech recognition systems to resolve ambiguity situations, and might also be used in optical character recognition, handwriting recognition, spelling correction, part-of-speech tagging, and machine translation.

"One striking clue to the importance of probabilities in language comes from the wealth of frequency effects that pervade language representation, processing, and language change. (...) Frequent words are recognized faster than infrequent words, and there is a bias toward interpreting ambiguous words in terms of their more frequent meanings. Frequent words lead leniting changes and are more prone to reduction in speech. Frequent combinations of phonemes and structures are perceived as more grammatical, or well formed, than infrequent combinations. The relative frequency of derived words and their bases affects the morphological decomposability of complex words" (Bod et al., 2003).

It is important to understand how frequency affects language processes and how the various aspects of the cognition processes are driven and self-organized by usage. Humans seem to track, record and exploit the occurrences of various kinds of events. This might make it fundamental to understand the statistical properties of language to better understand how a language works. Pierrehumbert (2003) proposes that linguistics constraints are a result of statistical robust generalizations that might be effectively learned, transmitted and exploited. The symbolic concept of a phoneme is a probabilistic distribution over a continuous phonetic space. The process of learning, recognition and classification of phonetic exemplars is a task of adjusting the phonemic membership functions. Under this point o view, language is a probabilistic process. Knowlegde of phonotatics involves knowledge of co-occurence probabilities of phonemes, and the well-formedness of a string of phonemes is just a combined product of the contributions of its subparts. "Such phonotactic probabilities are exploited in speech perception for segmentation, and they affect well-formedness judgments, influence pronunciation, and affect behavior in linguistic tasks such as creating blends" (Bod et al., 2003).

The effect and influence of probabilities on Language is present at different levels. Baayen (2003) presented the influence of the frequency of occurrence at the morpheme level, showing that the individual's choice among concurring affixes is strongly biased by the frequency of occurrence of these affixes. At the word level, the processing and representation of words is strongly influenced by lexical frequency and this behavior is independent of morphological composition of words. This frequency effect is manifested in ambiguity resolution, phoneme reduction, language change and speed of access (Bod et al., 2003). Individuals also track the co-occurrence of words, what influences speech

comprehension and production. Jurafsky (2003) argues that more frequent word pairs have a shorter processing time and may also suffer a phonetic reduction. Low-probability words (regarding the surroundings) are more likely to receive a pitch accent. Jurafsky (2003) provide evidence that "people track the probabilities of syntactic structures", what influences the processing time of sentences and structures, and it is also involved in disambiguation.

"Language displays all the hallmarks of a probabilistic system. Categories and well-formedness are gradient, and frequency effects are everywhere. We believe all evidence points to a probabilistic language faculty. Knowledge of language should be understood not as a minimal set of categorical rules or constraints, but as a (possibly redundant) set of gradient rules, which may be characterized by a statistical distribution" (Bod et al., 2003).

As remarked by Sedelow and Sedelow (1966), "the study of patterns formed in the process of the linguistic encoding of information, is of importance to any major research focusing upon or dependent upon the production or analysis of language". An increased interest on the statistical and numerical counts of linguistics objects is perceived on Languages research, where the *quantitative* aspects have gained an important status. The use of computational resource and the availability of digital information has made the quantitative analysis task possible, what was hitherto unfeasible.

The quantitative analysis is also the basis of *stylometrics* (quantitative stylistics) which deals with the analysis from the standpoint of individual or functional style. Style is regarded as a probabilistic concept by which selection (choice), conscious or unconscious, is responsible for creating a style, an emergent feature during the choice process in a universe of multiple alternatives for expressing an idea (Tuldava, 2004). The analysis of emergent patterns in communication and the influence on the linguistic style used is analyzed by Hancock et al. (2004) to investigate the truthful and deceptive dyadic communication. The studies show that "senders used more words overall, increased references to others, and used more sense-based descriptions (e.g., seeing, touching) when lying as compared to telling the truth. Receivers naïve to the deception manipulation produced more words and sense terms, and asked more questions with shorter sentences when they were being lied to than when they were being told the truth" (Hancock et al., 2004).

The first remarkable study in the statistics of languages was done by George Kingsley Zipf, linguistic professor of Harvard, during the 1920s and 1930s. Zipf and his students performed many word count experiments and determined that there is a relationship between the word's frequency of appearance in texts and its rank, the product of them is roughly a constant. That means, the distribution follows a power law: $f(k; s, N) = Ck^{-s}$, where f stands for frequency, C is a constant, k is the word rank, s the slope, the exponent

characterizing the distribution, and N is the number of elements in the set. Zipf's law is not, in fact, a law in the rigorous sense, but an empirical observation that has an apparent robustness.

The constant C is a normalizing constant that might be found by calculating the total number occurrences of all elements:

$$\sum_{k=1}^{N} f(k; s, N) = \sum_{n=1}^{N} Cn^{-s} \ . \tag{6.1}$$

Normalizing the occurrences of each element ranked k by the total of occurrences, leads us to

$$f(k; s, N) = \frac{k^{-s}}{\sum_{n=1}^{N} n^{-s}} \ , \tag{6.2}$$

that means, the constant C is given by

$$C = \frac{1}{\sum_{n=1}^{N} n^{-s}} \ . \tag{6.3}$$

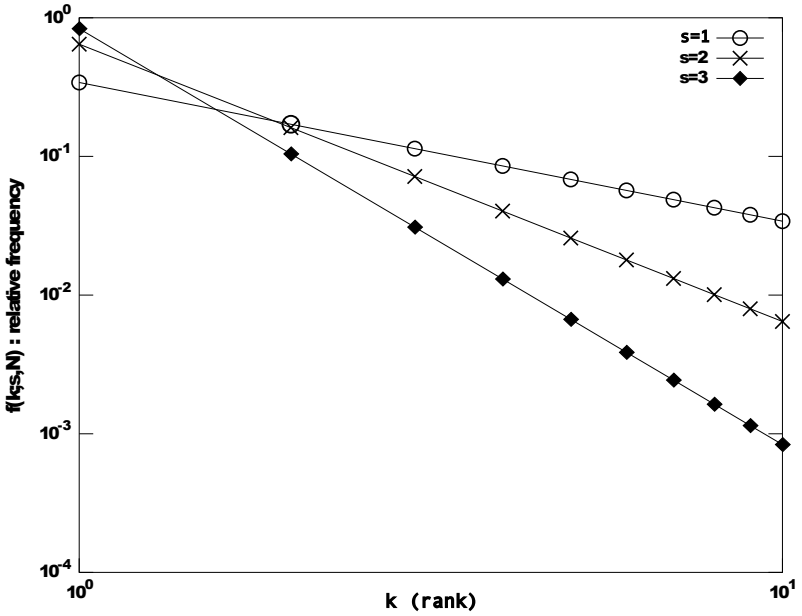


Figure 6.1: Zipf probability mass function for $N = 10$ on a log-log scale for different values of s.

Zipf developed the idea of using an intrinsic linguistic or psychological reason to explain this phenomenon observed in the world of words. He named his theory the 'Principle of Least Effort' to explain why frequently encountered words are chosen to be shorter in order to require a little mental and physical effort to recall them and utter/write them.

According to Alexander et al. (1998), Zipf's law seems to hold regardless the language observed. "Investigations with English, Latin, Greek, Dakota, Plains Cree, Nootka (an Eskimo language), speech of children at various ages, and some schizophrenic speech have all been seen to follow this law"(Alexander et al., 1998).

A study by Nowak et al. (2000), on the evolutionary dynamics languages, aim to understand how the transition from non-syntactic communications, typically found among many animals, could lead to syntactic communication, only found among humans. A model for the population dynamics of language evolution was proposed, and the conclusions show that "natural selection can only favour the emergence of syntax if the number of required signals exceeds a threshold value" (Nowak et al., 2000). The emergence of syntax might be responsible for Zipf's pattern that is observed in human communication.

Zipf's law[1] is also observed in other phenomena, for example: the magnitude of earthquakes (it is common to have many small earthquakes, but big ones are rare) (Abe and Suzuki, 2005); the population in cities (there are few megalopolises, but thousands of small cities) (Gabaix, 1999); the distribution of total liabilities[2] of bankrupted firms in high debt range (Fujiwara, 2004); the number of requests for web pages (Adamic and Huberman, 2002); etc.

In order to derive the statistics for the English language, data from the Gutenberg Project[3] database were collected. The top 100 most downloaded books were initially used as our analysis database. Perl scripting was used to read all 100 books, list the words and count their occurrences. The top 59 most frequent words in the database and their number of occurrence are listed below. Using this list, it is straightforward to create a log-log plot presenting the words' rank versus their frequencies of occurrence, which is shown in Figure 6.2.

1. the : 775911
2. and : 471916
3. of : 414499
4. to : 350613
5. a : 277321
6. in : 226505
7. i : 200689
8. that : 173083
9. he : 162183
10. it : 145364
11. was : 130804
12. his : 129300
13. you : 118473
14. with : 114122
15. is : 112640
16. for : 107245
17. as : 102009
18. not : 96636
19. be : 86896
20. but : 81643
21. had : 80327
22. at : 76688

[1]A cumulative distribution with a power-law form is said to follow a *Zipf's law* or a *Pareto distribution*. Zipf's law usually is used in a context to refer the relation between frequency of occurrence of an event relative to it's rank. Pareto's law is given in terms of the cumulative distribution (CDF), i.e. the number of events larger than a certain value is given by an inverse power of that value. A Power law is simply the probability distribution function (PDF) associated with the CDF given by Pareto's law.

[2]Liability, in financial accounting, "is a present obligation of the enterprise arising from past events, the settlement of which is expected to result in an outflow from the enterprise of resources embodying economic benefits" (definition of the International Accounting Standards Board, IASB).

[3]The Gutenberg Project (`http://www.gutenberg.org/`) is the oldest digital library and was founded in 1971 by Michael S. Hart. It has over 33,000 items in its collection and are free because their copyright has expired.

23. her : 75761
24. on : 75493
25. my : 73879
26. him : 72258
27. have : 68463
28. this : 67572
29. all : 65960
30. me : 64560
31. by : 63944
32. which : 63051
33. she : 57839
34. they : 57770
35. from : 56128
36. or : 52089
37. so : 51617
38. said : 50040
39. no : 48930
40. are : 45831
41. one : 43822
42. what : 41575
43. them : 41320
44. were : 40475
45. will : 39733
46. if : 38421
47. there : 38209
48. we : 37944
49. when : 37385
50. their : 36721
51. who : 36109
52. an : 35485
53. your : 33401
54. would : 32582
55. do : 31225
56. out : 30165
57. then : 29682
58. been : 29502
59. up : 28860

. . .

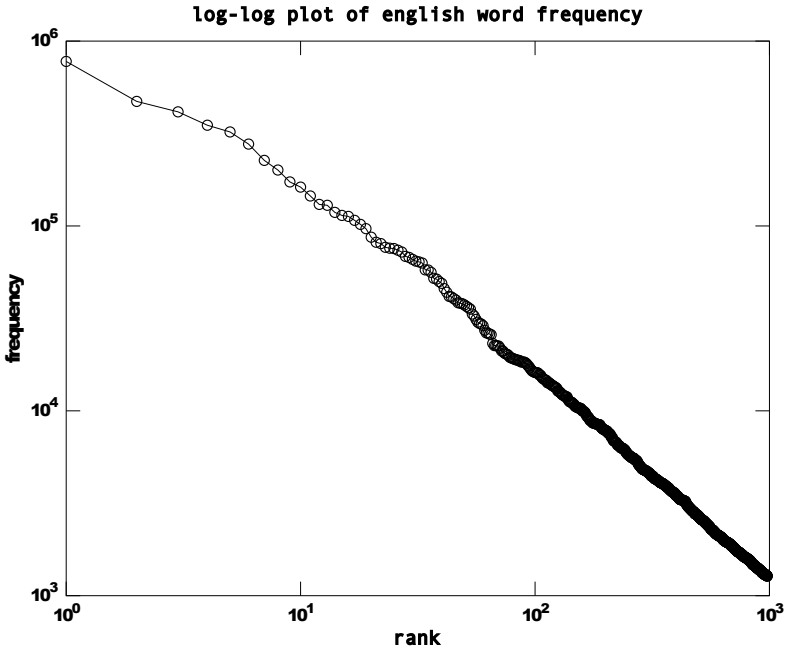


Figure 6.2: Log-log plot of words rank versus frequency of occurrence (only the 1,000 first words are presented).

In the analysis presented, all strings of letters separated by white spaces were considered words. The definition of word itself is, in certain aspects, dubious. It is considered as the smallest free form that can be uttered or written and carries a meaning. This is the concept introduced by Bloomfield (1926) but it is doubtful since some words are not minimal free forms, like *the* and *of*, which make no meaning by themselves. There are also those meanings that require two strings of letters to express themselves, like: *stock market*, *apple tree*, *carbon dioxide*, *electric guitar*, *hot tub*, *cotton candy*, *dental floss*, *hot dog*, among others. There are some words that are clearly compound ones, like: *newspaper*, *thumbnail*, *copperhead*, *eyelid*, *bedroom*, etc.; and from the previous example, we

see that there is a tendency of some two words becoming a compound word, for example, *hot dog* it also written as *hotdog*. In other languages, like German, it is even harder to settle the boundaries of words. In 1996, the German word *Donaudampfschifffahrtselektrizitätenhauptbetriebswerkbauunterbeamtengesellschaft* (Association for subordinate officials of the head office management of the Danube steamboat electrical services) was added to the Guinness Book of World Records as the largest word in that language. But the longest word that is not created artificially seems to be *Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz* (Cattle marking and beef labeling supervision duties delegation law).

Anyway, we are interested in language as a spoken means of communication, not a written one, although it is known that one influence the other. However we don't have a speech database and it would be hard and time consuming to create one, we adopted the Carnegie Mellon University (CMU) Pronouncing Dictionary to get the phonetic transcription of the words in our database. The CMU Dictionary is a public domain pronouncing dictionary for North American English that contains over 125,000 words and their transcriptions. It is used as the American lexicon for the Festival Speech Synthesis System and also for the CMU Sphinx speech recognition system. Our database extracted from the Gutenberg's books is made of 164,444 words (types) and 14,285,332 occurrences of such words (tokens). It is still small when compared to the number of entries of some American English dictionaries (Merriam-Webster has more than 470,000 entries), but it suffices for a first analysis. Unfortunately only 25% of the words matched the words in the CMU Dictionary, some of them because of spelling differences, like the one in *colour - color*, *favour - favor* and *neighbourhood - neighborhood*; other because they are old-English words, like *thyself*, *milady*, *beheld* and *picot*; and all the rest is attributed to misspelling, proper names, abbreviations, foreign words or even words that are really not part of the CMU Dictionary vocabulary.

Using the database, transcribed through the CMU Dictionary, it was possible to draw some conclusions from the phones usage in English. We present first the list of phones and their frequency of occurrence:

1. ə : 44539
2. t : 33131
3. n : 31928
4. ɪ : 28845
5. s : 21928
6. d : 20032
7. r : 18563
8. i : 16482
9. l : 15816
10. ɛ : 13896
11. m : 13072
12. ɝ : 12640
13. k : 12308
14. w : 11107
15. z : 10744
16. ð : 10720
17. v : 10407
18. h : 10009
19. f : 9391
20. ɑ : 8744
21. æ : 8635
22. b : 8390
23. u : 7972
24. p : 7501
25. ɔ : 7429
26. eɪ : 6196
27. aɪ : 6148
28. oʊ : 5283
29. ʃ : 4915
30. ŋ : 4861
31. g : 3351
32. tʃ : 2501
33. j : 2462
34. θ : 2309
35. ʊ : 2276
36. aʊ : 2242
37. dʒ : 2100
38. ɔɪ : 326
39. ʒ : 314

The data above are used to plot the frequency of occurrence of the phones against their rank, as seen in Figure 6.3, shown in a log-log plot. We may observe that the data don't form a straight line, what could be expected, since we are dealing with a very small set and Zipf's law is characteristic of a class of Large Number of Rare Events (LNRE) (Baayen, 2001). Since the number of distinctive phones is quite small, we will observe a reasonably well estimated phones' probabilities. As we move forward to larger units: bigrams, trigrams, etc., we expect to increase the chances of observing a power law relation.

Li (1992) argues that the pattern observed in Zipf's law has significant value since it is a natural observation on random processes. Even though, we observe that, as we analyze larger chunks of symbols, the relation between the frequency of occurrence of these chunks and their rank approximate progressively a Zipf's law (see Figure 6.4).

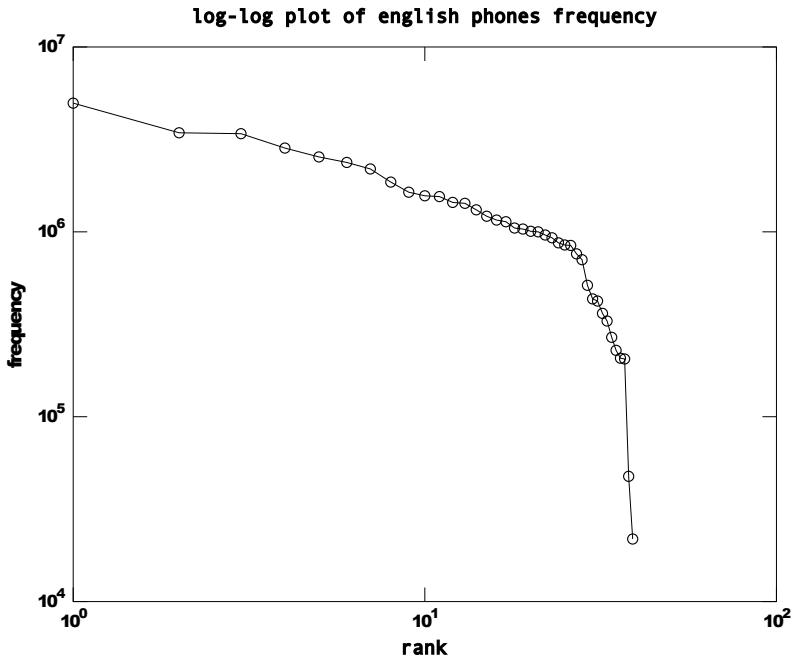


Figure 6.3: Log-log plot of phones rank versus frequency of occurrence.

Another way to see the Zipf's law is as a distribution of Pareto. The Pareto principle states that, for many events, most of the effects come from the minority of the causes. This principle was named after the Italian economist Vilfredo Pareto. In 1906, he observed that 80% of the land in Italy was owned by 20% of the population. The phrase "The kth most frequent word has n occurrences" may be stated, from the Pareto's perspective, as "k words occur n or more times". A Pareto plot is shown in Figure 6.5. It combines a bar chart displaying percentages of the English phones (categories) with a line graph showing cumulative percentages of these categories. We observe that the 8 first most frequent phones ([ə, t, n, s, ɪ, r, d, l]) account for half of all phones occurrences in the data. In

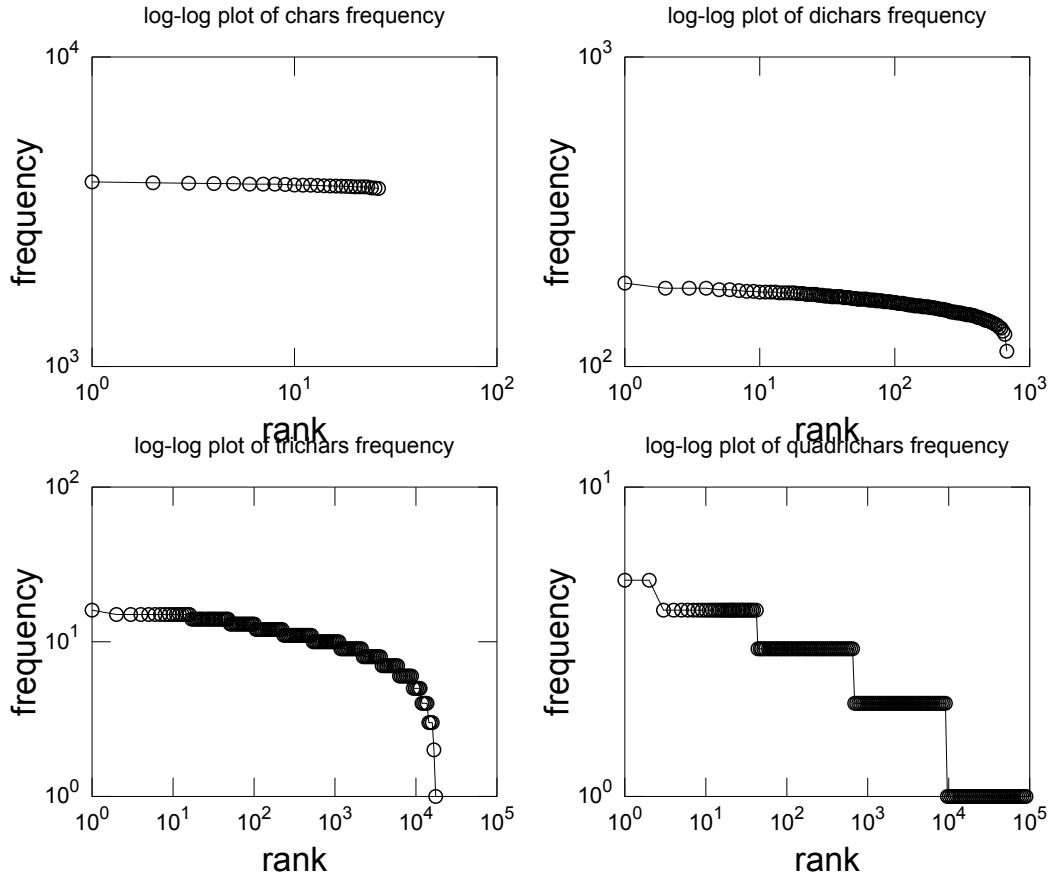


Figure 6.4: Log-log plot of the rank of chunks of different sizes versus their frequency of occurrence.

turn, the Figure 6.6 is a simple ordinary plot of the phones and their respective frequency of occurrence.

One important quantity that might be obtained from a database is a measure of the information being transmitted by a source using a given set of symbols. The notion of information is dealt by the concept of entropy (what be will discussed in more details in another section) and is usually measured in bits. The greater the entropy of a source, the greater is the uncertainty associated with its output and then it is also greater the amount of information encoded in its messages. The entropy of spoken English, calculated from the data above, gives 4.84 bits per phone. The entropy of written English was estimated by (Schneier, 1996), who found it to be between 1.0 and 1.5 bits per letter. It was also estimated by Shannon (1951), having found a value between 0.6 and 1.3 bits per letter, in a experiment where subjects were asked to predict the next letter in a English text. The relative entropy, Kullback–Leibler divergence, of English phones compared to a uniform random distribution is 0.45 bits.

Zipf believed that the change in frequency was responsible for triggering the mechanism

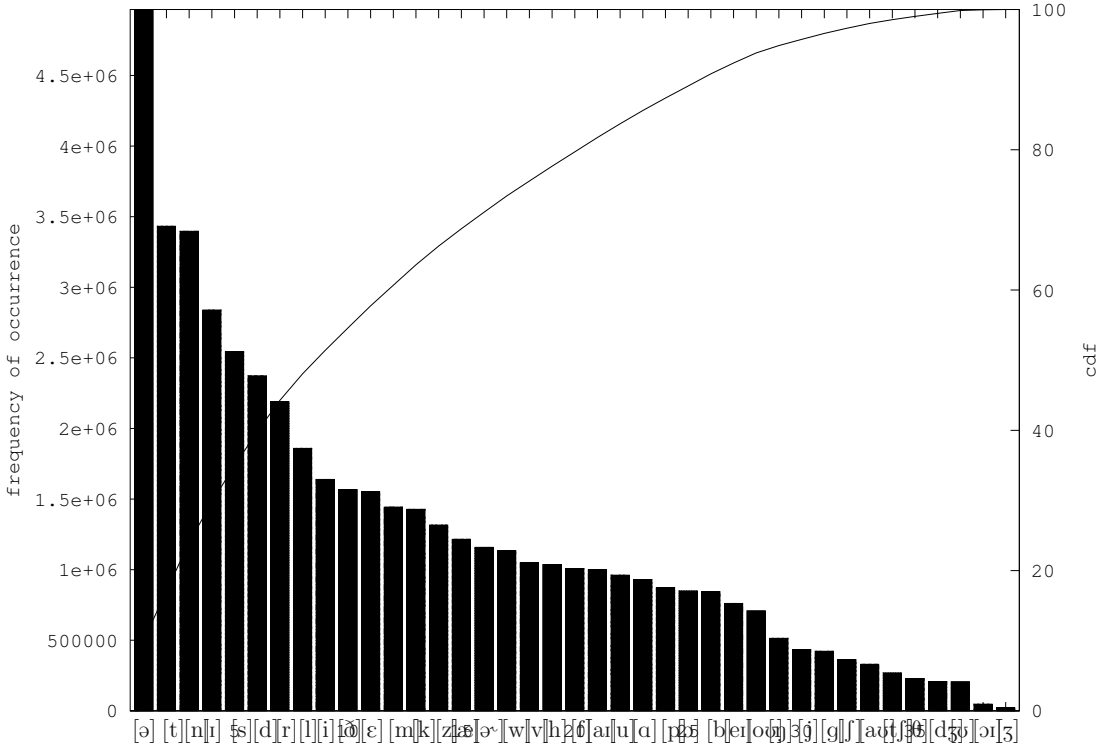


Figure 6.5: Pareto plot of the English phones.

of combination. He supposes that the frequency of occurrence of a lower level unity is proportional to the number of higher level structures it will appear on. The low specificity of a lower level unit makes it more suitable for been used in making higher level structures. Figure 6.7 presents the relationship between the frequency of occurrence of English phones and the number of English words they appear on.

The distribution of the occurrence of a given phone is not uniform across the rank of words in a language. Nor can't we conclude that the high frequency of certain phones is due to the high frequency of the words they appear. Figure 6.8 shows the probability of occurrence of phones across the rank of words. We might observe that the probability of occurrence of a given phone is greater in rare words then in usual ones. Since high rank words are longer, the probability of occurring any phone is greater than on low rank words.

Analyzing the sound systems of human languages, we conclude that they show remarkable regularities. Although humans are able to pronounce and perceive a great variety of speech sounds, languages use only a subset of them. Under another perspective, the categories created by languages, in order to achieve communication, form a small subset of possible categories of speech sounds that have a certain regularity across lan-

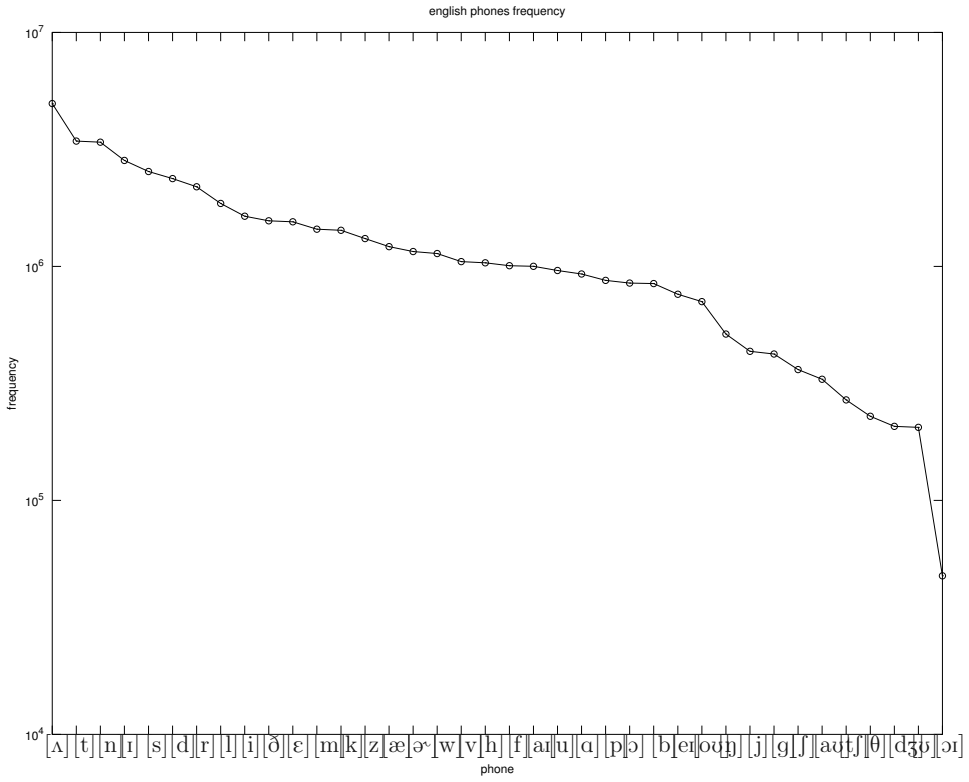


Figure 6.6: Frequency of occurrence of English phones.

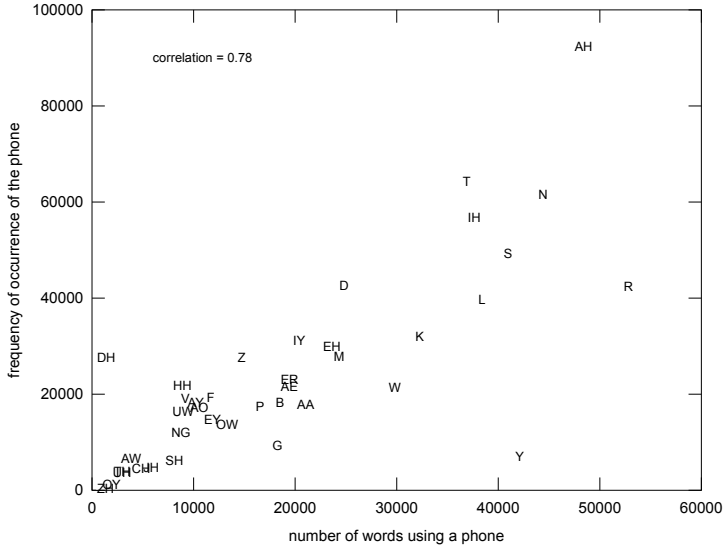
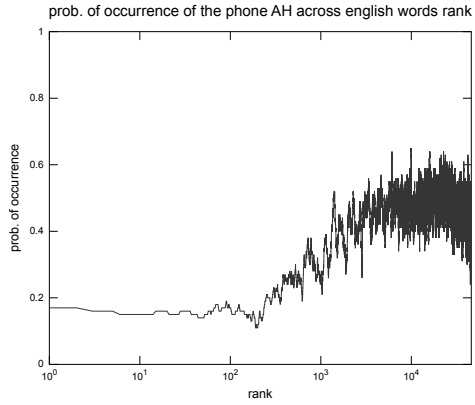
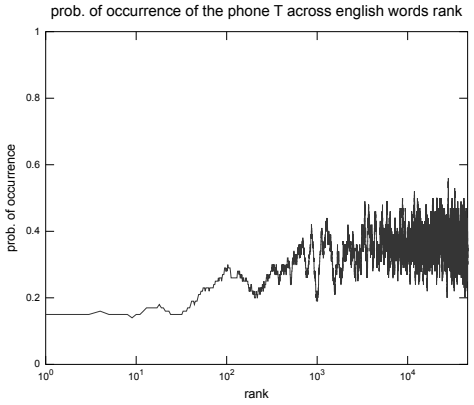


Figure 6.7: Relation between the frequency of occurrence of phones and the number of words they appear.

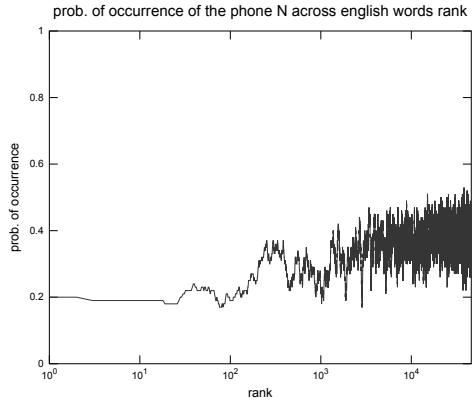
guages. Those regularities might be explained by innate human cognitive capacities or by functional constraints of communication. Such constraints would be created by the



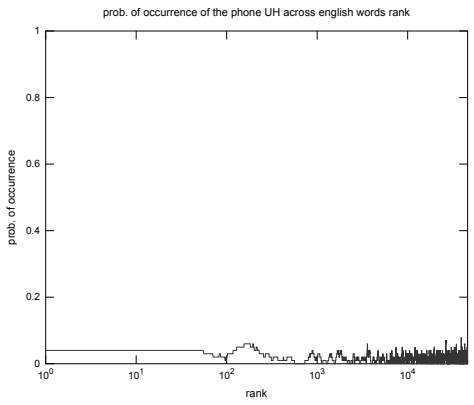
(a) Probability of occurrence of [ə] in words versus words rank.



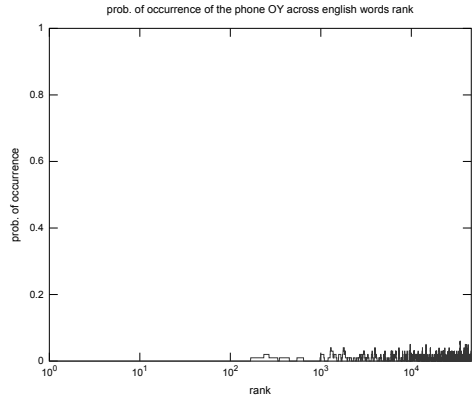
(b) Probability of occurrence of [t] in words versus words rank.



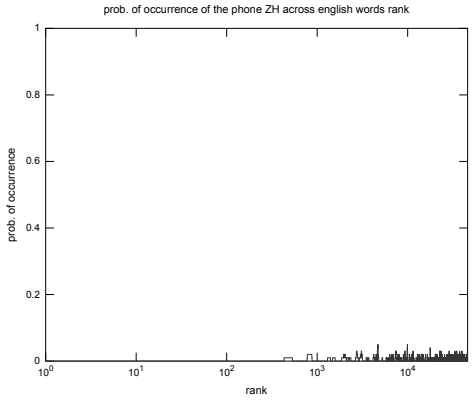
(c) Probability of occurrence of [n] in words versus words rank.



(d) Probability of occurrence of [ʊ] in words versus words rank.



(e) Probability of occurrence of [ɔɪ] in words versus words rank.



(f) Probability of occurrence of [ʒ] in words versus words rank.

Figure 6.8: Probability of occurrence of certain phones in words across the rank of words.

requirement of language to be a robust learnable means of communication, which would be responsible for the existence of redundancy, predictability, distinguishability and the usage of sounds of easy understandability and repeatability.

In the analysis of 451 languages of the world made by Maddieson (1984) using the UPSID (the UCLA Phonological Segment Inventory Database), 921 different speech sounds are found. As shown in Figure 4.1, most of the languages (66.3%) have a repertoire of 20 to 37 speech sounds. In each language, we use only a subset of all the speech sounds available (using as a reference the UPSID and the 919 phones on this database, a language typical uses between 1.2% and 15.3% of the available speech sounds). Analyzing the subsets used in each language, we present bellow a list of the 20 most frequent consonants and the 10 most frequent vowels (see tables 6.1 and 6.2).

Table 6.1: List of the 20 most frequent consonants in UPSID.

consonant	m	k	j	p	w	b	h	g
n. of languages	425	403	378	375	332	287	279	253
frequency	94.2	89.4	83.8	83.2	73.6	63.6	61.9	56.1
consonant	ŋ	ʔ	n	s	tʃ	ʃ	t	f
n. of languages	237	216	202	196	188	187	181	180
frequency	52.6	47.9	44.8	43.5	41.7	41.5	40.1	39.9
consonant	l	n̪	t̪	ɲ				
n. of languages	174	160	152	141				
frequency	38.6	35.5	33.7	31.3				

Table 6.2: List of the 10 most frequent vowels in UPSID.

consonant	i	a	u	ɛ	o/ɔ	e/ɛ	ɔ	o	e	a
n. of languages	393	392	369	186	181	169	162	131	124	83
frequency	87.1	86.9	81.8	41.2	40.1	37.5	35.9	29.0	27.5	18.4

Comparing tables 6.1 and 6.2, we observe that the most frequent vowels are not so frequent across languages in comparison to the most frequent consonants. In the UPSID, there are 180 vowels, 16 glides[4] and 725 consonants, which strongly outnumber the others. The languages in UPSID present from 6 to 117 consonants (and glides), averaging 22.7 and from 3 to 28 vowels, averaging 8.1. The ratio consonant to vowel ranged from 0.69 to 15.33, averaging about 3.38. On average, languages present three times as many consonants as vowels. Considering that the database with 451 languages has in its inventory 652 consonants and 269 vowels, we see that across the languages, the number of vowels used

[4]Glide or semivowel contrast with vowels by being non-syllabic, they functions as the syllable boundary rather than nucleus. In our analysis we will consider only the following as semivowels: palatal approximant, labial-palatal approximant, velar approximant, labial-velar approximant, and their possible variations.(Martínez-Celdrán, 2004)

corresponds from 0.9% to 17.9% (averaging 3.0%) of the possible vowels; and the number of consonants used goes from 0.9% to 14.6% (averaging 3.5%) of the consonants in the database. We might then conclude that the segments are chosen proportionally among all the possibilities, there is no tendency in choosing consonant over vowels, or the other way around. Analyzing these results and those on tables 6.1 and 6.2, we conclude that there is some sort of stronger constraint on consonants that push some of them to become more frequent across languages. A constraint of this kind also exists for vowels, but it is much weaker. We also observe that, in general, languages use a speech repertoire where the number of consonants is greater than the number of vowels. There are only 14 languages (Kashmiri, Bruu, Dan, Klao, Kaingang, Apinaye, Barasano, Yagua, Cubeo, Japreria, Panare, Andoke, Maxakali, Vanimo) in which the number of vowels is greater or equal to the number of consonants. There are 427 speech sounds that are present in only one language. The group of sounds that appear in 10 or fewer of the 451 languages in the database sum up more that 80% of all 919 sounds in the database.

Figure 6.9 presents the relationship observed across languages of the usage of frequent or rare segments with the number of segments used in that language. In order to do so, a frequency index is proposed by Reetz (2010). This index is the average of the segment frequencies of the segments in a language. Each segment in the UPSID has a segment frequency that is the number of languages that contain a specific segment divided by the number of languages in UPSID. Each language has a certain segment-repertoire. The frequency index of a language is calculated as the average of the frequency index of all segments in a language repertoire. "If a language has only few segments, it is likely that these are rather common in the languages in UPSID. On the other hand, a language with many segments will also have many segments that are uncommon in the UPSID database"(Reetz, 2010).

Using the UPSID database, we might find what speech segments co-occurs with each other in different languages. When addressing the UPSID database, we shall use the term *co-occurrence* to make reference to phones that occur in the same inventory. This remark is important in order to avoid a possible confusion with the usual meaning of *co-occurrence*, that is used to assign phones that are neighbours in an utterance. We may find what are the most frequent co-occurring segment for another given segment (some results are in Figure 6.10 and Table 6.3). We notice that there is a large number of co-occurring segments for each one taken as a reference. The most frequent segments across languages have a great number of co-occurring segments, which represent approximately from 30% to 60% of all segments in UPSID. On the other hand, the most infrequent phones in the database have just a few co-occurring segments, around 3%. Observing the graphics in Figure 6.10, we see that, for each reference segment, the co-occurring ones has a rapidly

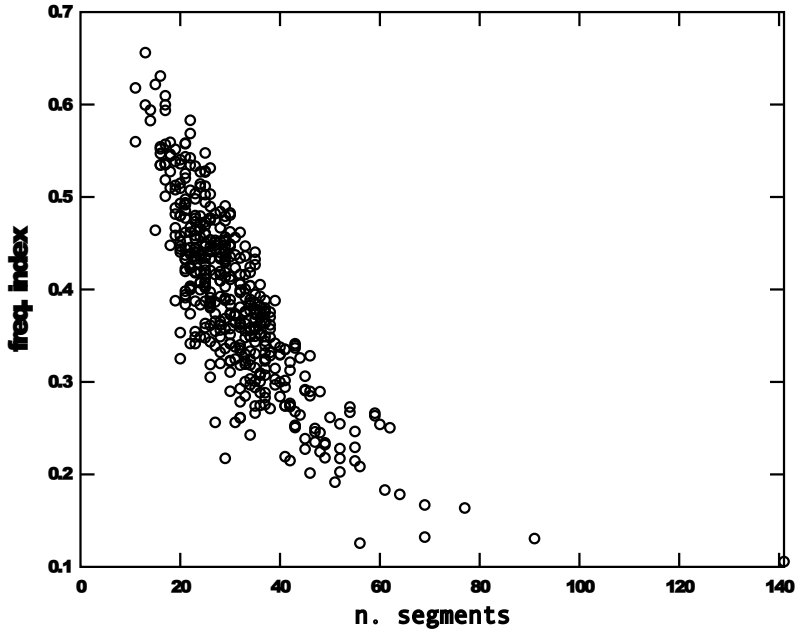
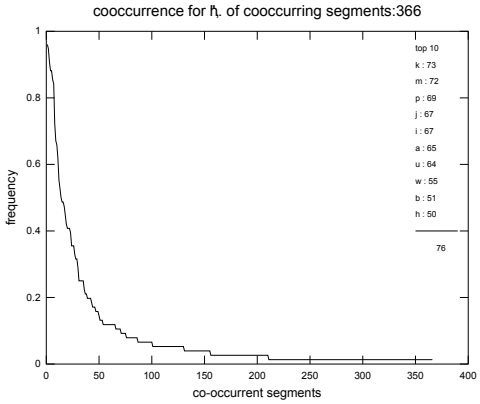


Figure 6.9: Relation between the frequency index and the number of phones in a language. (Data from UPSID)

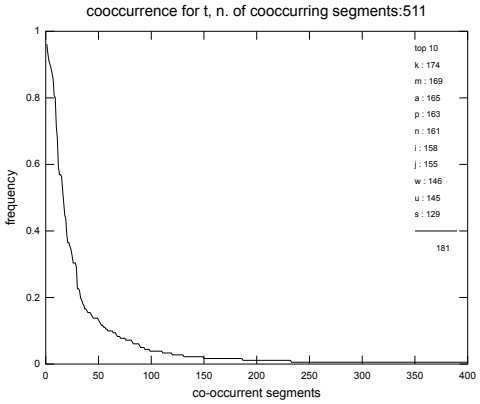
decreasing frequency of co-occurrence, what is expected from a random distribution. The high co-occurrence of certain pairs, in some cases, might be explained by the features they share. As an example, if we consider the bilabial consonants, we see that, as we take one as reference, the others appear in the top-10 list, showing a strong adhesion of one bilabial consonant to another. Observing now the voiced alveolar sibilant fricative [z], its voiceless counterpart always appear, but doing the other way around analysis, taking [s] as a reference, its voiced counterpart [z] has a relative frequency of co-occurrence of 31.6%, figuring as the 29th in the list. Analyzing other pairs like [t]-[d], [k]-[g] and [p]-[b], it seems that the existence of the voiced counterpart subjects the existence of the voiceless much more emphatically than the other way around.

If we believe that speech symbols repertoires are not chosen randomly, we might wonder what guides the choices of a repertoire. Not only the way those symbols are arranged, but also the way they are used (combined) is important in the process of understanding choices. Using the Gutenberg's database transcribed with the CMU pronouncing dictionary, as described above, we might get information on phone clusters frequency of occurrence (see also Figure 6.11):

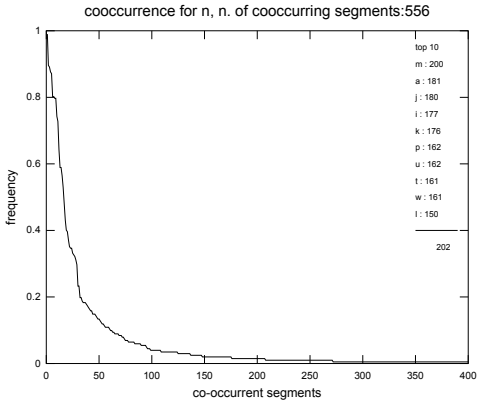
1. ən : 1296408
2. ðə : 785354
3. nd : 784028
4. st : 651129
5. əv : 489267
6. ɔr : 472069
7. ɪn : 470069
8. tu : 425544
9. tə : 420825
10. ɪŋ : 387096
11. ɪz : 380357
12. əl : 372847
13. ɛn : 349905
14. nt : 337193
15. æt : 284460
16. wɪ : 282526
17. ət : 265396
18. ɛr : 264293
19. rə : 261447
20. ɪt : 260544



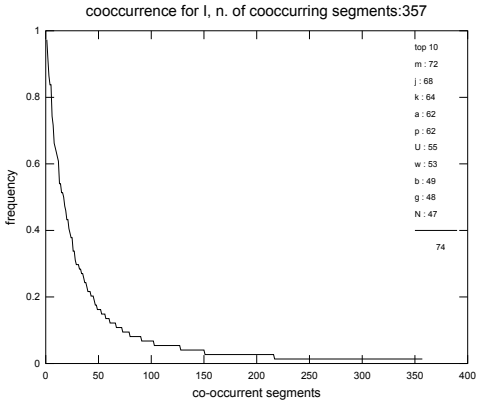
(a) This graph presents the co-occurrence frequency for phones in relation to [ə]. The phone [k] has a frequency of 96.0%. [m] follows with 94.7% and [p] with 90.7%. 39.8% of the phones in UPSID are co-occurring with the phone [ə].



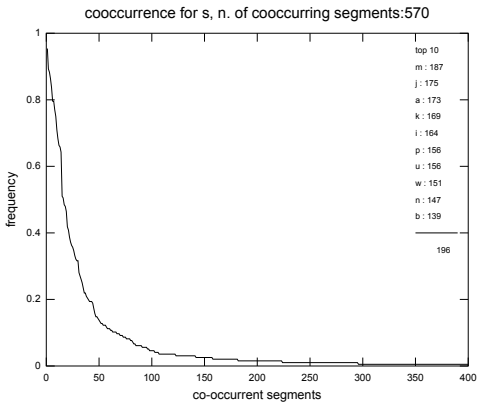
(b) This graph presents the co-occurrence frequency for phones in relation to [t]. The phone [k] has a frequency of 96.1%. [m] follows with 93.3% and [a] with 91.2%. 55.6% of the phones in UPSID are co-occurring with the phone [t].



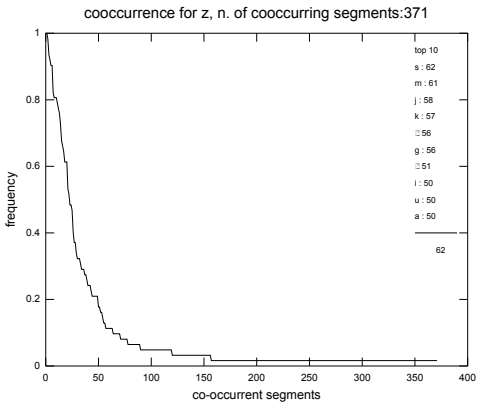
(c) This graph presents the co-occurrence frequency for phones in relation to [n]. The phone [m] has a frequency of 99.0%. [a] follows with 89.6% and [j] with 89.1%. 60.5% of the phones in UPSID are co-occurring with the phone [n].



(d) This graph presents the co-occurrence frequency for phones in relation to [l]. The phone [m] has a frequency of 97.3%. [j] follows with 91.9% and [k] with 86.5%. 38.8% of the phones in UPSID are co-occurring with the phone [l].



(e) This graph presents the co-occurrence frequency for phones in relation to [s]. The phone [m] has a frequency of 95.4%. [j] follows with 89.3% and [a] with 88.3%. 62.0% of the phones in UPSID are co-occurring with the phone [s].



(f) This graph presents the co-occurrence frequency for phones in relation to [z]. The phone [s] has a frequency of 100%. [m] follows with 98.4% and [j] with 93.5%. 40.4% of the phones in UPSID are co-occurring with the phone [z].

Figure 6.10: The frequency co-occurrence plots above are derived from the UPSID.

Table 6.3: List of phones and their top 10 co-occurring pairs with their relative frequency of occurrence (data from UPSID).

phone	co-occurring phone with their respective relative frequency (%)									
ə	k : 96.1	m : 94.7	p : 90.8	j : 88.2	i : 88.2	a : 85.5	u : 84.2	w : 72.4	b : 67.1	h : 65.8
t	k : 96.1	m : 93.4	a : 91.2	p : 90.1	n : 89.0	i : 87.3	j : 85.6	w : 80.7	u : 80.1	s : 71.3
n	m : 99.0	a : 89.6	j : 89.1	i : 87.6	k : 87.1	p : 80.2	u : 80.2	t : 79.7	w : 79.7	l : 74.3
ɪ	m : 97.3	j : 91.9	k : 86.5	a : 83.8	p : 83.8	ʊ : 74.3	w : 71.6	b : 66.2	g : 64.9	ŋ : 63.5
s	m : 95.4	j : 89.3	a : 88.3	k : 86.2	i : 83.7	p : 79.6	u : 79.6	w : 77.0	n : 75.0	b : 70.9
z	s : 100.0	m : 98.4	j : 93.5	k : 91.9	b : 90.3	g : 90.3	p : 82.3	i : 80.6	u : 80.6	a : 80.6
d	b : 96.7	m : 94.2	i : 91.7	a : 90.8	j : 90.0	n : 89.2	g : 87.5	u : 86.7	t : 85.8	k : 84.2
l	m : 98.9	j : 89.1	k : 86.8	n : 86.2	a : 86.2	i : 85.6	p : 81.0	w : 79.9	u : 79.9	s : 72.4
i	m : 93.9	u : 91.6	k : 89.6	a : 89.1	p : 82.7	j : 82.7	w : 73.8	b : 65.1	h : 62.6	g : 57.5
d̪	b : 97.5	m : 96.2	g : 93.8	j : 88.8	k : 85.0	t̪ : 83.8	i : 80.0	p : 76.2	u : 72.5	a : 70.0
m	k : 89.2	i : 86.8	a : 86.6	j : 85.2	p : 82.6	u : 81.6	w : 74.1	b : 64.2	h : 61.6	g : 56.7
n	m : 99.0	a : 89.6	j : 89.1	i : 87.6	k : 87.1	p : 80.2	u : 80.2	t : 79.7	w : 79.7	l : 74.3
k	m : 94.0	p : 91.3	i : 87.3	a : 86.6	j : 83.4	u : 82.1	w : 73.2	b : 62.8	h : 61.0	g : 54.1
g	b : 96.4	m : 95.3	i : 89.3	j : 87.4	k : 86.2	u : 84.2	a : 83.0	p : 76.7	w : 72.3	h : 63.6
p	k : 98.1	m : 93.6	a : 87.2	i : 86.7	j : 82.9	u : 81.3	w : 71.7	b : 60.5	h : 60.3	ŋ : 54.4
b	m : 95.1	i : 89.2	k : 88.2	j : 86.8	g : 85.0	u : 84.3	a : 84.3	p : 79.1	w : 72.5	h : 65.9
ʃ	m : 94.7	j : 91.4	k : 88.8	i : 84.0	a : 82.9	p : 80.2	u : 78.1	w : 74.3	h : 73.3	b : 68.4
ʒ	ʃ : 95.1	m : 95.1	j : 90.2	k : 90.2	b : 82.0	g : 80.3	i : 78.7	p : 78.7	u : 77.0	a : 70.5

21. əs : 246996
22. ju : 245715
23. hæ : 237724
24. hɪ : 237490
25. əm : 233793
26. ri : 219997
27. li : 219652
28. ɑn : 213755
29. æn : 213169
30. sə : 210385
31. ɪs : 208619
32. ðæ : 194602
33. əf : 193570
34. ɑr : 193525

⋮

1120 uaɪ : 1
1121 ɔɪɑ : 1
1122 bʃ : 1
1123 pv : 1
1124 iʊ : 1
1125 ɛoʊ : 1

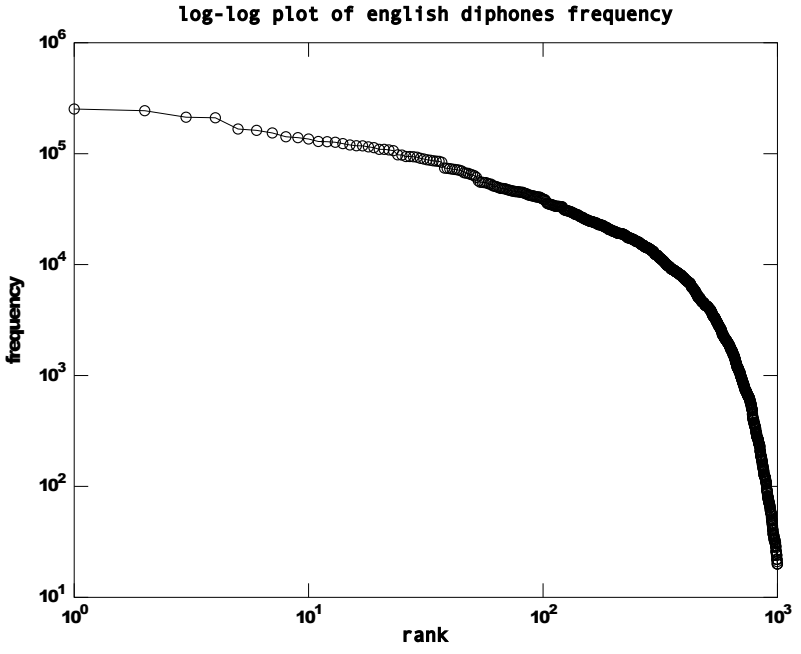


Figure 6.11: Log-log plot of the diphones frequency of occurrence versus their rank.

Normalizing the frequency of occurrence of each diphone by the frequency of occurrence of each phone that is part of them, we get a different ordering that is displayed in the list bellow, and illustrated by the log-log plot in Figure 6.12.

1. ʒə	5. dʒə	9. ju	13. ətʃ	17. əp	1122 aıh
2. ðə	6. bə	10. əm	14. nd	18. ⋮	1123 ɛoʊ
3. əv	7. əl	11. ɔr	15. kə	1120 zs	1124 ddʒ
4. ʃə	8. əf	12. əb	16. əg	1121 uaı	1125 tv

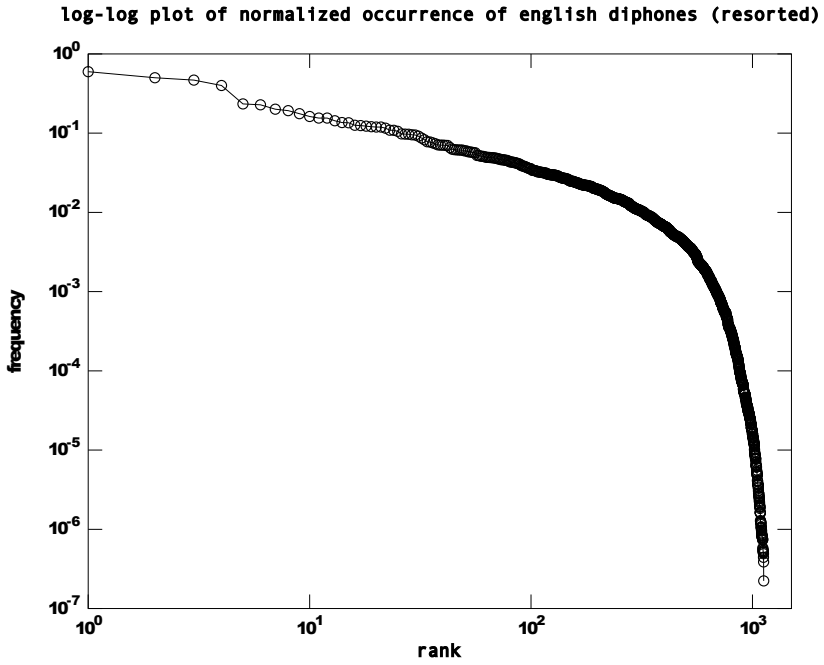


Figure 6.12: Log-log plot of the diphones normalized frequency of occurrence versus their rank. The normalization is made using the frequency of occurrence of each phone in the pair.

Using the information on the occurrence of diphones, it is possible to estimate the probability of occurrence of a subsequent phone for each previous occurring phone. Those probabilities were computed and are displayed in Figure 6.13 as a matrix. Each entry (i, j) in the matrix refers to the probability of occurrence of phone j followed by phone i. The phones in the first position of a diphone (prior phones) are arranged along the y axis of the figure, and the phones in the second position (posterior phone) of the diphone are displayed along the x axis. The phones are arranged according to the their frequency of occurrence in the language. We observe in the figure that the most frequent phones (in the left part of the figure) also have an average higher conditional probability of occurrence regardless which the prior phone is.

Another analysis here presented consists of computing the number of elements (letters and phones) used to build up words. Figure 6.14a shows the number of occurrence of

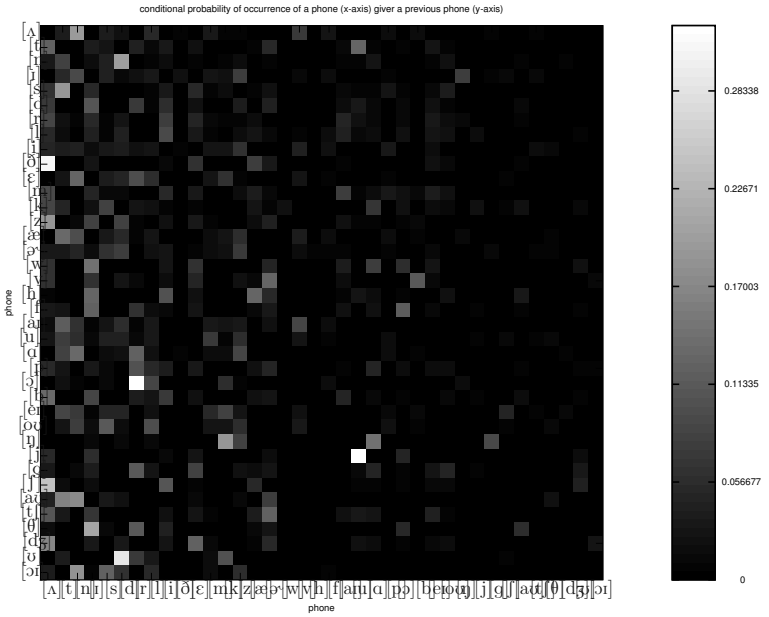
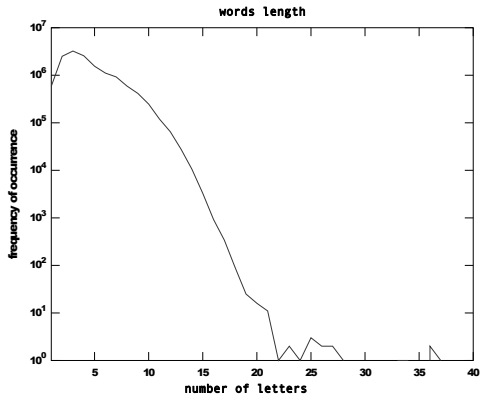
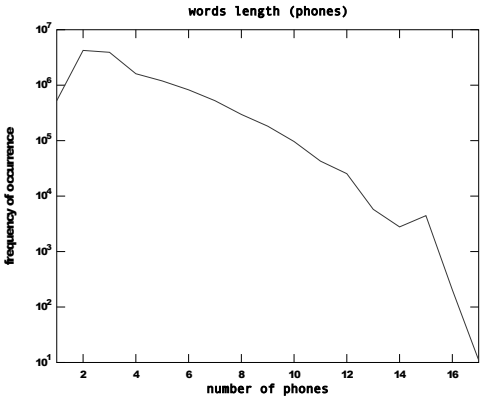


Figure 6.13: Probability of occurrence of a phone given another previous phone.

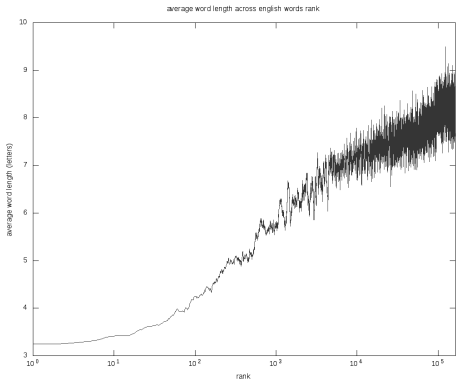
words with a certain letter-length. We observe that the peak occurs in 3 letter words. Figure 6.14b is a similar graphic, displaying the number of occurrence of words with a certain phone-length. The peak appears in two phone words. For every L symbol word made up of a combination of symbols taken from a set of N elements, there are N^L possible combinations. The last two graphics 6.14c and 6.14d show the average word length across word rank, showing that the most frequent words, on average, have a short length; the unusual words are, on average, significantly longer.



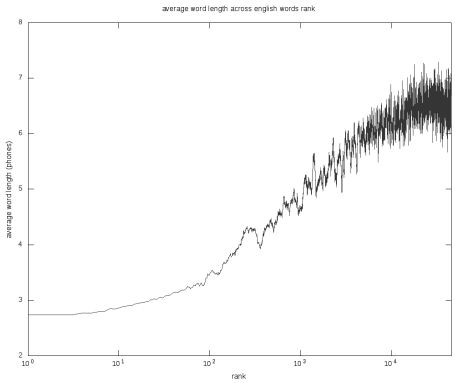
(a) Frequency of occurrence of words of a given length (letters).



(b) Frequency of occurrence of words of a given length (phones).

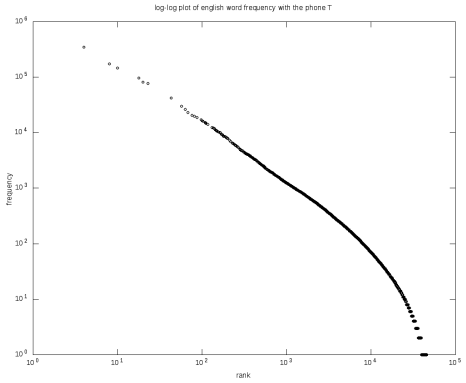


(c) Average word length (letters) across word rank.

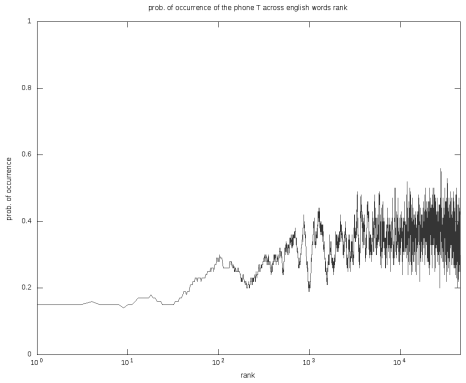


(d) Average word length (phones) across word rank.

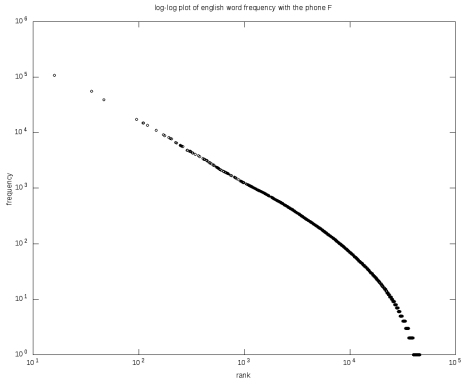
Figure 6.14: Words length statistics (letters and phones) and how it does deviate from a simply random combination of symbols.



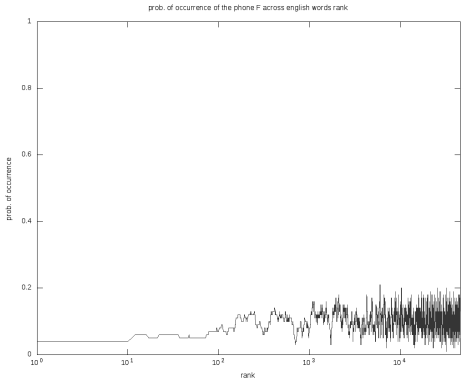
(a) Log-log plot of the frequency of occurrence of words with phone [t] versus their rank.



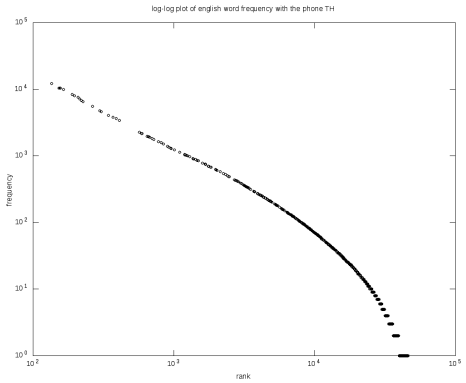
(b) Probability of occurrence of words with phone [t] versus the rank of words.



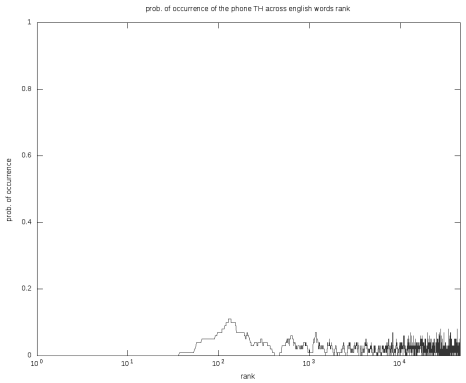
(c) Log-log plot of the frequency of occurrence of words with the phone [f] versus their rank.



(d) Probability of occurrence of words with phone [f] versus the rank of words

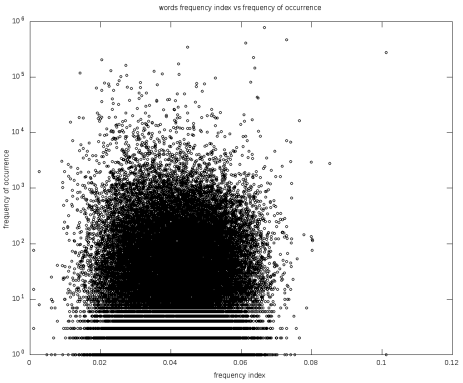


(e) Log-log plot of the frequency of occurrence of words with phone [θ] versus their rank.

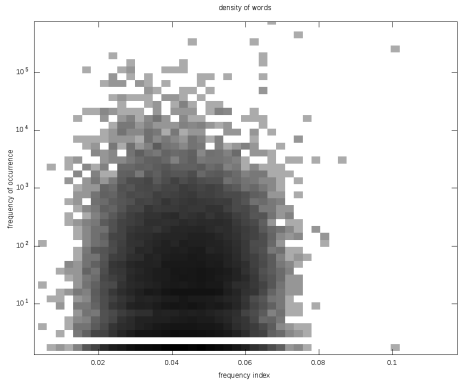


(f) Probability of occurrence of words with phone [θ] versus the rank of words

Figure 6.15: Two types of graphics are presented to verify the contribution of word frequencies to the final phone frequencies. The left plot shows the occurrence of words with a certain phone. The right one shows an estimation of the probability of occurrence of a certain phone versus the rank of words.



(a) Each spot represents a word with a given frequency of occurrence and a frequency index. For a better visualization the frequency of occurrence is displayed in a logarithm scale.



(b) Density of words in each partition on the frequency of occurrence vs. frequency index space. The largest number is displayed in black and it refers to 578 words. White represents no word found in a spot. The gray scale is displayed in a logarithm fashion for better visualization.

Figure 6.16: Relationship of word frequency of occurrence and word frequency index. The frequency index is the average probability of occurrence of the phones that make up the word.

7

Artificial and Natural Language Compared

> Running texts deal necessarily with certain particular subject matter, and cannot therefore be regarded without further ado as random samples of the vocabulary of the language. The question therefor arises, what to do in order to satisfy the condition of a sensible application of statistics also on the vocabulary level.
>
> Gustav Herdan (1966)

In this chapter we analyze the statistical properties of natural English language and three different random types of computer generated texts. We use the conventional Zipf analysis and the inverse procedure which will be further explained. We also use the Shannon entropy to characterize the language source.

We have seen that many things that might be measured in Nature have a typical size or 'scale', but other things vary over an enormous dynamic range and we observe that they follow a behavior that is described by a Zipf law. Power-law distribution occur in a wide range of different phenomena, for example, city populations, size of earthquakes, size of moon craters, solar flares, computer files, number of hits on web pages, and human

language, among many others. Random generated texts also exhibit this Zipfian behavior, and we shall here analyze and compare with Natural Language in order to establish the statistical properties that distinguish them.

7.1 Language Patterns

Language is a complex adaptive system driven by its usage. It is important to understand how patterns emerge by usage and what are their influence on language evolution over time. Recent experiments show that it is a human behavior to track patterns and occurrences even on an artificial grammar (Saffran et al., 1996b, 1999; Saffran and Wilson, 2003). These observations support the hypothesis that the patterns in a language cause great influence on the cognitive representations of this language.

The idea of a usage-based theory is that use creates patterns and structures. This premise is well applied into the emergence of a grammar: some patterns become frequently used, turning into conventions, or fossilized grammatical patterns (Givón, 1979; Hopper and Thompson, 1980, 1984). It is all a process of repetition and ritualization, that is responsible for the emergence of new structures. This process of ritualization is also observed in the establishment of grammatical phonotactical patterns. Speakers make their judgments of the grammaticality of phonotactic patterns based on the frequency of co-occurrence of certain consonants and vowels in a language (Frisch, 1996; Pierrehumbert, 1994).

There are evidences that words and frequent phrases are units of lexical storage and manipulation (Bybee, 2006). As such, there is no reason to treat them differently from other mental records of a person's experience. Most psycholinguistic models include word frequency as an important component of speech perception, learning and usage. Nowadays these models have gone more general and included all sorts of probabilistic information about words, phrases, and other linguistic structures represented in the mind of a language user. According to these models, the frequency of usage plays a role in language comprehension, production and learning (Jurafsky, 1996; MacDonald, 1993; Gregory et al., 1999; Brent and Cartwright, 1996).

In order to study some patterns in English Language and inquire to what extent the patterns and structures that emerge are just a result of chance, we propose here to follow Zipf's steps and to analyze the text *Ulysses* by James Joyce and compare it with different types of artificially random generated texts, ranging from a white random process to a Markov model with transitions probabilities like the ones presented in *Ulysses*.

7.2 Zipf Revisited

"Zipf's law may be one of the most enigmatic and controversial regularities known in linguistics. It has been alternatively billed as the hallmark of complex systems and dismissed as a mere artifact of data presentation. Simplicity of its formulation, experimental universality and robustness starkly contrast with obscurity of its meaning"(Manin, 2008).

The first analysis made by Zipf was based on the data of James Joyce's *Ulysses*. The book has 260,430 running words and, considering the work was done in the 40s, it already had a large size, creating difficulty on the analysis process. Dr. M. Joos has carefully extracted quantitative information from these 260,430 running words, concluding that there are 29,899 different words in it. He did also ranked those words in the decreasing order of their frequency of occurrence.

From the Gutenberg database we could download a text-only copy of *Ulysses* and count the occurrence of words within it. There were a total of 29,165 different words found in the 271,848 running words. The slight differences found might be due to different editions used by each analysis (the Gutenberg's version is based on the pre-1923 print editions), due to different assumptions used to assign what is a word (in our analysis we don't consider compound words with spaces in between, nor compound words with hyphens in between), what explains the smaller number of observed types and the greater number of tokens in our analysis. An ordered list of these words follows bellow, presenting the 29 most frequent words in Joyce's *Ulysses* and their respective frequency of occurrence:

1. 15,126 : the
2. 8,256 : of
3. 7,284 : and
4. 6,582 : a
5. 5,043 : to
6. 5,004 : in
7. 4,226 : he
8. 3,333 : his
9. 3,009 : i
10. 2,840 : s
11. 2,795 : that
12. 2,560 : with
13. 2,528 : it
14. 2,134 : was
15. 2,126 : on
16. 2,083 : you
17. 1,962 : for
18. 1,786 : her
19. 1,526 : him
20. 1,461 : is
21. 1,341 : all
22. 1,303 : at
23. 1,289 : by
24. 1,208 : said
25. 1,198 : as
26. 1,189 : she
27. 1,103 : from
28. 1,053 : they
29. 1,036 : or

...

A relationship between the rank of words and their frequency of occurrence might be observed: their product is roughly a constant. This is better visualized in the log-log graphic of the rank versus frequency of occurrence. Figure 7.1 depicts this relation. The continuous line presents the behavior found in *Ulysses*, where an almost straight line

appears, showing the intrinsic relation in the data.

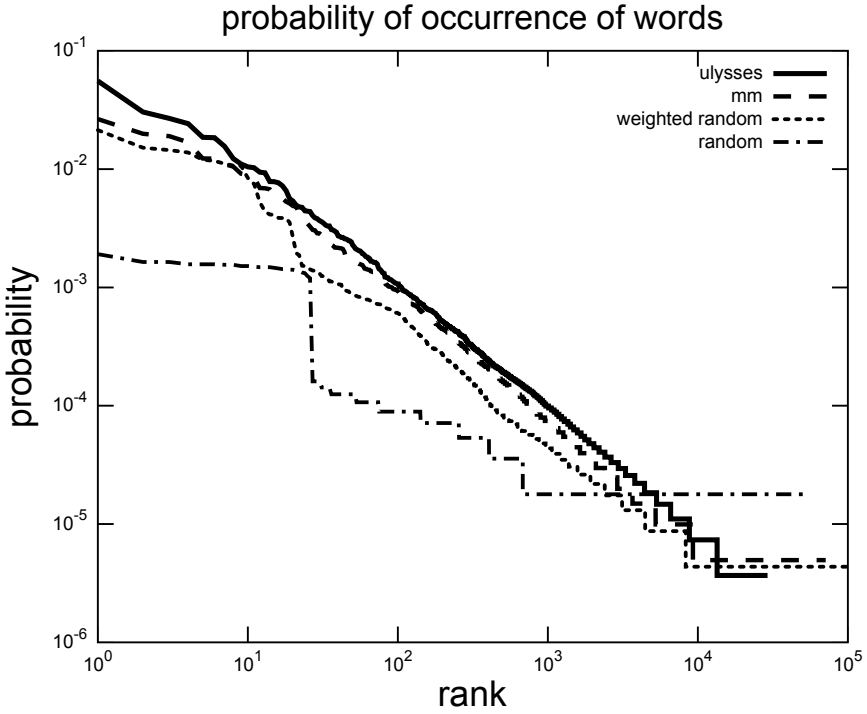


Figure 7.1: The rank-frequency distribution of (pseudo-)words in James Joyce's *Ulysses* and random texts using the phones/diphones probabilities. The random curve presents the frequency of occurrence of pseudowords derived by text created by a white process and having the same length as *Ulysses*. The weighted random curve is created by randomly choosing symbols with the same probabilities as the phones in *Ulysses*. The *mm* curve presents the result of a random text derived by a Markov Model where the transition between phones has the same probabilities of the transitions found in *Ulysses*.

Zipf also presents the comparison curve of R. C. Eldridge data, which consists of 43,989 running words, with 6,002 different words, of combined samples from American newspapers. The concordance on the curves "clearly show that the selection and usage of words is a matter of fundamental regularity of some sort of an underlying governing principle that is not inconsistent with our theoretical expectations of a vocabulary balance as a result of the Forces of Unification and Diversification" Zipf (1949).

The Zipf curves, as presented in Figure 7.1, must inherently be monotonically decreasing curves, since the data is ordered and ranked according to the frequency of occurrence. Li (1992) points that "Zipf's law is not a deep law in natural language as one might first have thought. It is very much related to the particular representation one chooses, i.e., rank as the independent variable". If Zipf's law arises at randomly generated texts with no linguistic structure, we might conclude that the law may be a statistical artifact rather than a meaningful linguistic property.

Random generated sequences of letters results in the formation of letter chunks. Those

letter chunks might be sorted according to their frequency of occurrence and, by doing that, chunks of the same length will present approximately the same frequency, which will create a stair case pattern as presented by Li (1992). The decay follows approximately a Zipf's law slope ($\sim 1/r$), a power law with exponent 1 (one). The results presented in Figure 7.1 shows that white random text follows clearly a different pattern compared to the natural text, but as the random generator process assumes certain characteristics, it might have a pattern closer to the one found in natural texts.

The simple model considered by Li (1992) is due to Miller (1957); Mandelbrot (1965) and it is known as *random typing* or *intermittent silence* model. It is simply a random generator of characters, where a certain symbol is designed as a word-delimiting character. As seen here, the words formed by this approach show a Zipffian-like frequency behavior, but the number of different words of a certain length is exponential in length, what diverges from what is observed in a natural language, since it is in fact not even monotonic. The *intermittent silence* model might not be used to draw a conclusion that the Zipf's law is linguistically shallow (Mandelbrot, 1982), since that model is "shallow" itself.

Considering that the probability of occurrence of phones in a random generated text is not uniform, we experience different results that distance from the white random case and get closer to the natural text. Figure 7.1 presents three different randomly generated texts: 1) equal probability phones, white noise (depicted by dashed-point line); 2) weighted probabilities, using the same probabilities of the phones encountered in *Ulysses* (small dashed line); 3) text generated by a Markov Chain using the states' transitions probabilities equal to the probabilities in *Ulysses* (big dashed line).

In a natural language not all combinations of speech structures are possible, and they don't happen with the same probability. We might observe this when comparing the number of existing words for a given length. When a random generation process is used, the number of different words (types) will be greater than the number of words in a natural language, and that will be valid for every word length. What we observe is that a natural language presents a higher probability of occurrence for small words. In our compared example, the highest probability occurs on words of length 4 and 3. This behavior is quite different from the random pattern, as we might observe in Figure 7.2.

It is important to note that when we consider the random generated texts, we are generating a string of symbols that we are regarding as phones, and for that reason we consider a set with the same size of the set of phones. This assumption is made because we are assuming that phones are our unity of analysis. When we are dealing with natural text, we are using a phonetic transcription dictionary, as previously stated.

The examples of random generated text in Figure 7.2 show that when we consider a weighted probability of occurrence for the randomly generated symbols, we create the

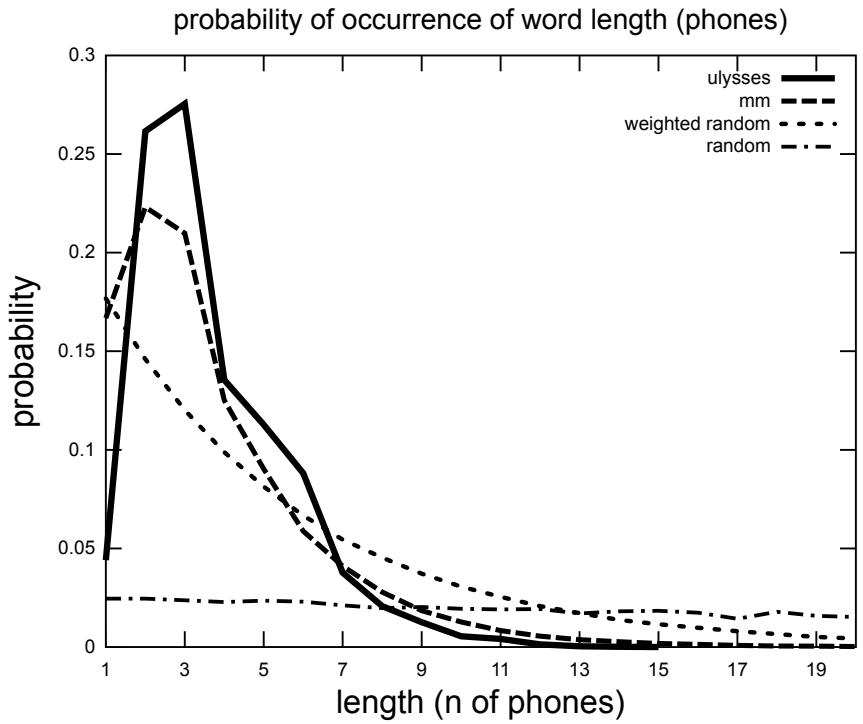


Figure 7.2: Compared plot of the frequency of occurrence of (pseudo-)words length in *Ulysses* and random texts as described in the text.

effect of making the shorted pseudo-words more probable, but what we observe in a natural language is different, we don't see a monotonically decreasing probability, but rather a peak on words of length 4. This effect is slightly achieved when we consider a Markov Model with transitions probabilities equal to the probabilities found in the natural text. To explain the behavior observed in Figure 7.2, we might consider the role of three forces: 1) a human cognition ability that makes us assign shortened codes to more frequent signs; this effect causes a tendency in the probability to monotonically decay as the length of words increases; 2) the number of possible symbols chunks that might be created, regarding that the symbols probabilities are not uniform, this will also creates a monotonically decreasing probability pattern; 3) there are restrictions imposed on certain combinations of phones, provoking a greater effect on short words, due to the smaller amount of possible combinations of symbols. We could also hypothesize that there are restrictions on the sequence of phones that overcome the immediate vicinity of that phone, and this effect would be stronger on phones that are closer, and weaker on phones that are far apart, also causing the smaller probabilities on very short chunks, where this restriction poses a stronger force, and being overcome on longer ones.

7.2.1 The Analysis of Smaller Units

In the previous analysis we focused on words as our speech units. A word is considered the smallest free form that can be uttered or written and carries a meaning. This is the concept introduced by Bloomfield (1926), but it is doubtful, since some words are not minimal free forms, like *the* and *of*, which carry no meaning by themselves. The status of words in a language is still a topic under debate and with no clear answers. "Words are units at the boundary between morphology and syntax serving important functions as carriers of both semantic and syntactic information and as such are subject to typological variation. In some languages words seem to be more clearly delimited and more stable than in others. The structural make-up of words depends on typological characteristics of languages" (Coulmas, 2003).

The main concern in Phonology is the study of language, as a means of spoken communication, that is built as a system created by sounds and structures. Under the phonological assumption, the concept of sound is not restricted at the physical level of speech realization, but also extends to symbolic level, where *sounds* are cognitive abstractions that provide the means for the representation of a language.

The Swiss linguist Ferdinand de Saussure is responsible for shifting the way linguistics was done and established a breakpoint with his posthumous publication *Course in General Linguistics* in 1916. Although many years have passed, the central concept of relating sound to meaning in a structured way has remained the same. In Saussure's model, the linguistic sign has three aspects: physical (sound waves), physiological (audition and phonation) and psychological (sounds as abstract units, which he calls 'sound images').

The concept of *sound images* was represented by the phoneme, regarded as a linguistic distinctive unity which groups different sounds into single and not superposed categories. Each language has its own phoneme inventory, which is created by what are the possible distinctions made between speech sounds in each language. For example, English makes no distinction between aspirated and unaspirated sounds, but in Hindi such distinction exists (Ladefoged and Maddieson, 1996). Implicit in the idea of phoneme is that language may be segmented into a sequence of consecutive symbols in time. Apostel et al. (1957) showed that it is an important requirement in order to make speech a comprehensible communication process in most situations, especially under heavy corruption. This discrete aspect of language guarantees the discreteness at all other levels of the language analysis, what is the basis of our linguistic knowledge.

The term *phoneme* was proposed by the linguist Dufriche-Desgenettes as a substitute for the German *Sprachlaut* (*speech sound*) in the early 1870s (Dresher, 2011). It had, by that time, the meaning of what we now call *speech sound* or *phone*. The concept of *phoneme* changed with Saussure, who used it "to refer to a hypothesized sound in a proto-

language together with its reflexes in the daughter languages" and further the meaning was recast by Kruszewski, bringing the synchronic notion, that we have today, of a set of alternating elements that were interpreted by Baudouin as an invariant psychophonetic element that may be realized in different forms and as proposed by Jones, a family of sounds that, for practical purposes, are accounted as the same (Dresher, 2011). Today the phoneme no longer holds a central place in phonology theory, but it has not disappeared from phonological theory nor its demise has come. Many evidences have challenged the phoneme status and its definition and properties (Port, 2007; Port and Leary, 2005; Port, 2006), which suggest that the theory of phonology is not yet completely defined.

"The point of concern at present, however, is not to devise a system of symbols whereby the sequence of sub-gestures constituting a speech-sound can be noted, but rather to remark that speech-sounds and phonemes may be viewed as constellations, or configurations, of articulatory sub-gestures, arranged partly or completely in sequential order, some sequences running concurrently with others (e.g. the occlusion is concurrent with the voicing of *d*). Although there is not a single speech-sound, or variant form, of a phoneme in any language which cannot be conceived of as a constellation or configuration of the type envisaged above, yet a complete and accurate description of even a single speech-sound in terms of sequences is practically impossible. However, by conceiving phonemes as constellations of this order, we have found a very useful method of comparing a number of specific phoneme pairs which have essential sequences in common"(Zipf, 1949).

The units of language (phonemes, syllables and words) are well represented in a writing system, for the process of writing requires the dissection of the speech stream into its constituents distinguishable parts. The process of dissection happens in different levels, so that the assemblage of its parts creates meaning on the speaker discourse. "Every writing system maps onto a linguistic system, it embodies and visibly exhibits the dissection of units of language and thus linguistic analysis"(Coulmas, 2003). It is reasonable then to investigate the patterns and structures of a language through its written counterpart.

In order to perform such analysis, we used the text *Ulysses* from Joyce, and created a transcription from written text to phones sequence. The Carnegie Mellon University (CMU) Pronouncing Dictionary (Weide, 2008) was used to get the phonetic transcription of each word in our database, according to the General American (GA) English, which is the major accent of American English. In the CMU Pronouncing Dictionary, words are coded by the phonetic transcription code *Arpabet*, where each code is a distinct sequence of ASCII characters. GNU tools, Python and Octave scripts were used to process and analyze the data.

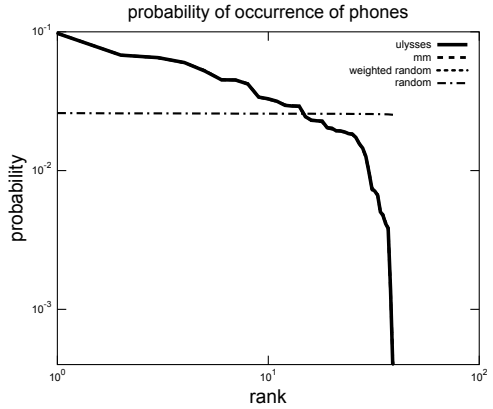
7.2.2 Phones, Diphones and Triphones

In this section we present the results observed as we draw Zipf plots of the phones, diphones and triphones, comparing the natural text with the random generated texts, as previously described. Figure 7.3a presents the compared probability of occurrence of phones. As all random texts were generated using the phone probabilities or transition probabilities from *Ulysses*, they presented the same behavior, only the white random process shows a different result, since all phones are equiprobable. There are 39 phones, and from a visual inspection of Figure 7.3a we might conclude that the Zipf's law does not hold for this very small set of symbols, but instead a logarithmic relation exist, as also pointed by Kanter and Kessler (1995). As we analyze higher order structures, we shall notice that it will progressively approach a Zipf's relation as we go from phones to diphones, then from diphones to triphones, until we reach the words' level, which clearly presents a Zipf's law.

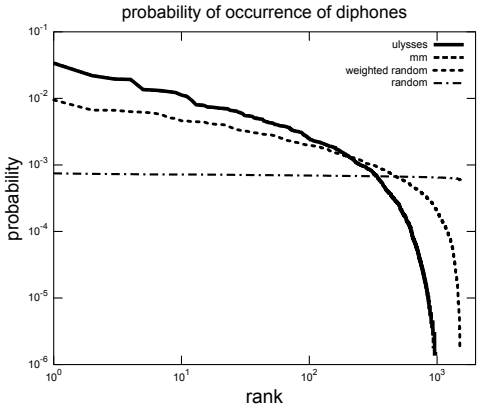
It is interesting to observe how phones occur in a language and the way they relate with one another creating higher order structures under different constraints. The acquisition of a language is based on tracking those multiple regularities in the input stimuli. Only if we analyze long data it will be possible to observe emergent regularities, and the more complex the regularities under observation are, the more data will be needed to establish a distinction among those regular patterns. Those regularities are known as grammatical structure of a language, and since early ages, infants can perceive the difference of grammatical and ungrammatical sentences (Saffran and Wilson, 2003).

The plot in Figure 7.3b shows the probability of occurrence of diphones found in the same texts described before. The white random process still preserves the equiprobability on the occurrence of diphones. The Markov Model still follows very closely our reference text, and our zero order model does not present the same characteristic as the reference. When we take into consideration transition probabilities as our source model, we observe that the formation of some diphones is facilitated, while the formation of others is hardened. The behavior of phones clearly doesn't agree with a power law relation, but not, observing the behavior of diphones we start to see a Zipf's law being created, as a result of the increase on the length of our symbol set.

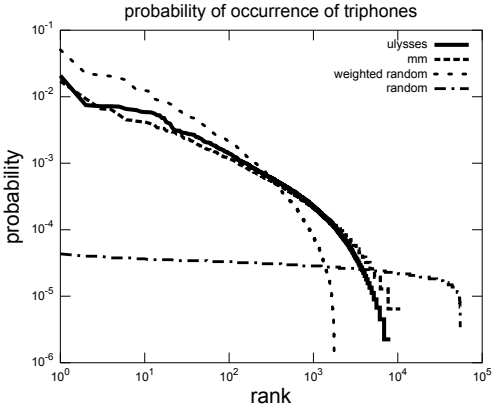
As we move forward from phones, diphones and triphones to larger chunks, what we may observe is that the behavior gets closer to the Zipf law observed when we analyzed words as our structural blocks. Observing the compared plot stated in Figure 7.3c we might observe that only the Markov model follows closely the natural behavior of triphones. The frequency of occurrence, or predictability, of a speech unit is believed to be related to its complexity when regarding acoustic realization (Jurafsky et al., 2001), and the complexity is further associated with the duration of the given structure. Us-



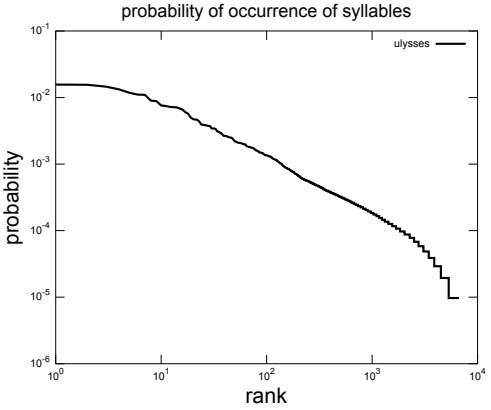
(a) Comparison of phone frequency of occurrence. All three curves present the same pattern.



(b) Comparison of diphone frequencies of occurrence. The behaviour of diphone frequencies generated by the white and weighted random text diverges from the others.



(c) Comparison of triphone frequencies of occurrence. The behaviour of triphone frequencies generated by the Markov Model does not follow closely the pattern presented in *Ulysses* as was still observed with diphones.



(d) Frequency of occurrence of syllables in *Ulysses*. The syllabification was done computationally using online dictionaries.

Figure 7.3: Compared frequency of occurrence plot for phones, diphones and triphones in *Ulysses* and in random texts generated as previously described, a white random source; a weighted random source with phones probabilities equal to the one found in *Ulysses*; and a Markov model with transition probabilities between phones as the probabilities in *Ulysses*.

ing annotated speech corpora, Jurafsky et al. (2001) concluded that when regarding the frequency of occurrence of n-phones as a function of its duration, the relation is an inverted U-shape curve, so that short n-phones have smaller probability of occurring than middle-length n-phones and longer n-phones also have smaller probability. Comparing the frequency of occurrence of words for a given length we observed different results as we regarded phones or syllables as our building units for the words. As previously shown in Figure 7.2, short words have smaller probabilities than middle length words, and the peak occurred in 3 phone long words. When the length of a word is regarded as the number of syllables, then we observed a monotonically decaying curve of probability of occurrence of words versus word length. Although there is not a clear straight relation between the number of syllables in a word and its duration, from our observations in Section 7.3, it seems reasonable to infer that the duration is proportional, and that would lead again to the conclusion that, on higher order analysis, the behavior observed in the frequency of occurrence of words is different from its constituent parts. The U-shape curve is reshaped into a straight line as syllables or phones are combined to build up words.

It seems fascinating that languages and other natural phenomena, all of them a result of different complex systems, share a certain property which is described by a single simple law. The same patterns are also observed when analyzing words and n-phones in different languages. Miller (1965) proposes two different approaches to explain the ubiquitous observation of Zipf's law: "Faced with this massive statistical regularity, you have two alternatives. Either you can assume that it reflects some universal property of human mind, or you can assume that it reflects some necessary consequence of the laws of probabilities. Zipf chose the synthetic hypothesis and searched for a principle of least effort that would explain the apparent equilibrium between uniformity and diversity in our use of words. Most others who were subsequently attracted to the problems chose the analytic hypothesis and searched for a probabilistic explanation. Now, thirty years later, it seems clear that the others were right. Zipf's curves are merely one way to express a necessary consequence of regarding a message source as a stochastic process".

Larger units of speech are used as recognition units to model the coarticulatory effects, for example, syllables and triphones. The last one is quite often used in phonemic based speech recognition, for the context of a given phone is established by its preceding and following neighbors. The more features subsequent phonemes share, easier will be the articulation of this sequence, since the transitions tend to be smoother. On the other hand, it is important to create contrast between adjacent phonemes, so that one phone is distinguishable enough from the phone next to it, what will enhance discriminability. This tradeoff might be perceived on the triphones usage profile. Figure 7.4 presents the probability of occurrence of triphones given their intradistances. For each triphone, two

distances are calculated: the distance from the second to the first phone; and the distance from the last to the middle phone. The distance measure used is the number of distinctive features not shared by two phones under comparison, as is explained in Chapter 8. The number of occurrences of each triphone occurring in *Ulysses* was added up according to its intradistances and the final probability of occurrence for each intradistance pair was calculated. We might observe in Figure 7.4 that the most occurring triphones are those with medium intradistances, which appear as a peak. That makes evident the tradeoff, previously argued, between articulatory ease and contrast. Triphones with small and large intradistances are seldom used, and some intradistance pairs (26 in the total, which represent 11.5% of the 225 possibilities) are not used at all.

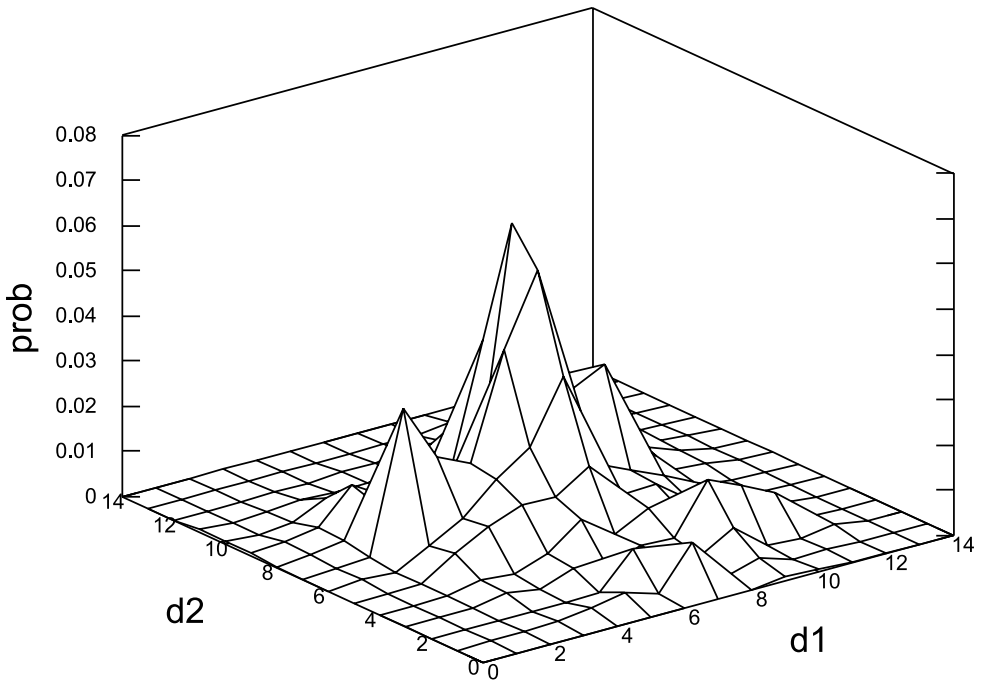


Figure 7.4: Given the intradistances in a triphone (the distance from the second to the first phone, d_1, and the distance from the third to the second phone, d_2), the probability of occurring triphones in *Ulysses* is presented. Triphones privileges medium intradistances, having a peak when the intradistances are around 8. The distance function used the number of non-shared distinctive features.

Figure 7.4 shows that there is a strong relationship between adjacent phones in a triphone structure, where the inter-distance between phones tend to an average value, and extreme values are not usual. It is natural then to speculate what would be this kind of relation within words. Figure 7.5 provides an analysis of the inter-phone distances within words. It presents, in a logarithmic scale, the frequency of occurrence of words for a given relation between the average inter-phone distances and the standard deviation of the inter-phone distances. Once again, the distance metric used is the number of distinctive features not shared by two phones. Among the plotted data, we observe

no tendency of increased occurrence of words for a certain relation between average and standard deviation of inter-phone distances. The plot of the number of words for a certain relationship of their inter-phone distances present a similar pattern, and normalizing the frequency of occurrence chart by chart with the number of words, we conclude that there is no aid making some inter-phone distance relation more probable than others. What we might observe from our chart and analysis is that there is a relationship between the average value and the standard deviation value for the inter-phone distances: when the average distance decreases, the deviation increases as a compensatory effect, for words have to maintain a certain phonemic variability within. The correlation coefficients (Pearson product-moment correlation coefficient) between the mean value and the standard deviation observed in the inter-phone distances within words is -0.67 and p-values of 0.01 (the probability that the observed data would have arisen if the null hypothesis, no correlation, were true).

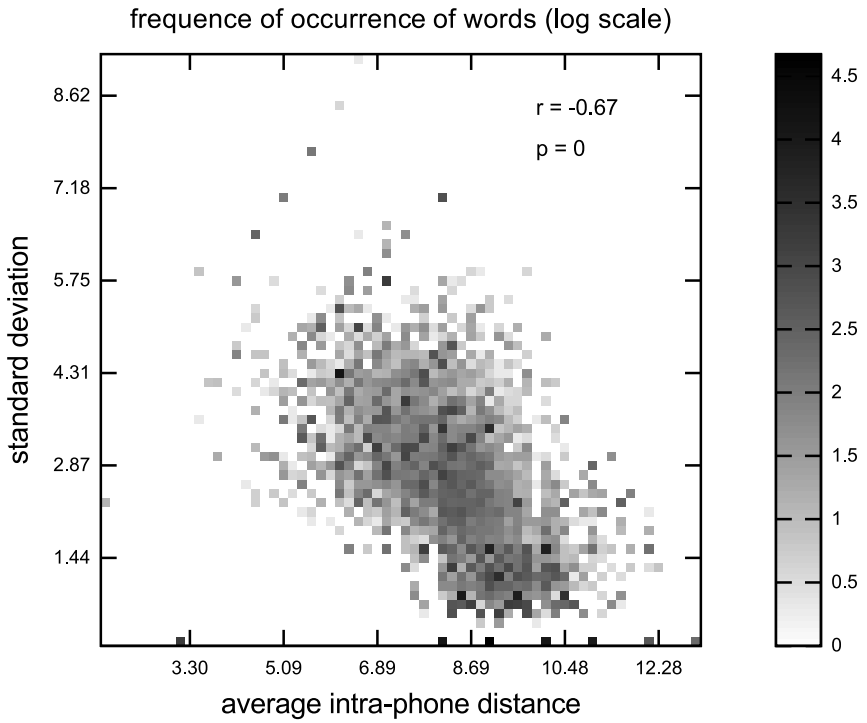


Figure 7.5: The words frequency of occurrence in *Ulysses* is presented in a logarithmic scale as a function of the average intra-phone distances and the standard deviation of these distances. The relationship between the number of existing words as a function of average and standard deviation of distances follows a similar pattern and our analysis doesn't show any effect that frequency has on this pattern. What we might observe is the tradeoff between average intra-phone distance and standard deviation, as the mean distance value decreases, the deviation increases as a compensatory effect. The correlation coefficient observed in this data (with 19,150 samples) is -0.67.

In order to analyze the variability of triphones that may be formed by a certain phone,

when regarding the transitions of distinctive features in a triphone, we considered the difference of the distinctive feature vectors between the middle phone (reference) and the first and last phones. As each reference phone occurs in many triphones combinations, we create a weighted sum of all these difference feature vectors, where the weight factor is the probability of occurrence of each triphone. For each phone we have built two vectors: one from the differences with the previous phone and the other from the differences with the next phone. Each of these vectors is represented by a column in the image presented in Figure 7.6. It is presented as shades of gray, when the gray level approaches white, then there is more variability (the difference is greater) and when it approaches black, there is less variability (the difference is smaller). We have ordered the features sorted by the accumulated variability across all different triphones for all phones in the set. The cumulative curve is plotted on the right. We also ordered the phones by the variability observed on their triphones. Each pair of columns in the image presented in Figure 7.6 corresponds to one phone that is transcribed in the axis bellow. The bottom curve in Figure 7.6 presents the variability for each phone, displaying the differences with the previous phone (left) and the differences with the next phone (right) in a triphone set. We might conclude that the features presented are sorted by their distinctiveness capabilities, that is: *syllabic*, *consonantal*, *sonorant*, *coronal*, *back*, *front*, *continuant acoustic*, *voice*, *high*, *strident*, *labial*, *approximant*, *nasal*, *distributed*, *dorsal*, *tense*, *lateral*, *labiodental*, *delayed release*, *spread glottis*, *trill*, *tap* and *constricted glottis*.

From the bottom plot in Figure 7.6 we might conclude that the variability between the first and the middle phones in a triphone is always greater than the variability between the middle and the last phones. Since every two side-by-side point corresponds to one certain phone, and they are always arranged in the order mentioned before (1st-2nd phones, 2nd-3rd phones), we observe that within each pair the curve presents a negative slope, and for that reason we might conclude that the first phone pair (1st-2nd phones) has a greater distinctive feature variability in every triphone. As we analyze larger chunks, we verify that this decreasing behavior on the average distance between phones does not hold anymore.

Preliminary studies show the existence of consistent frequency distribution patterns in other languages. These regularities are believed to be a result of language as a stochastic process and a product of usage and self-organization. If the underlying principle of organization of languages is the same, we are supposed to observe similar patterns when analyzing language statistics.

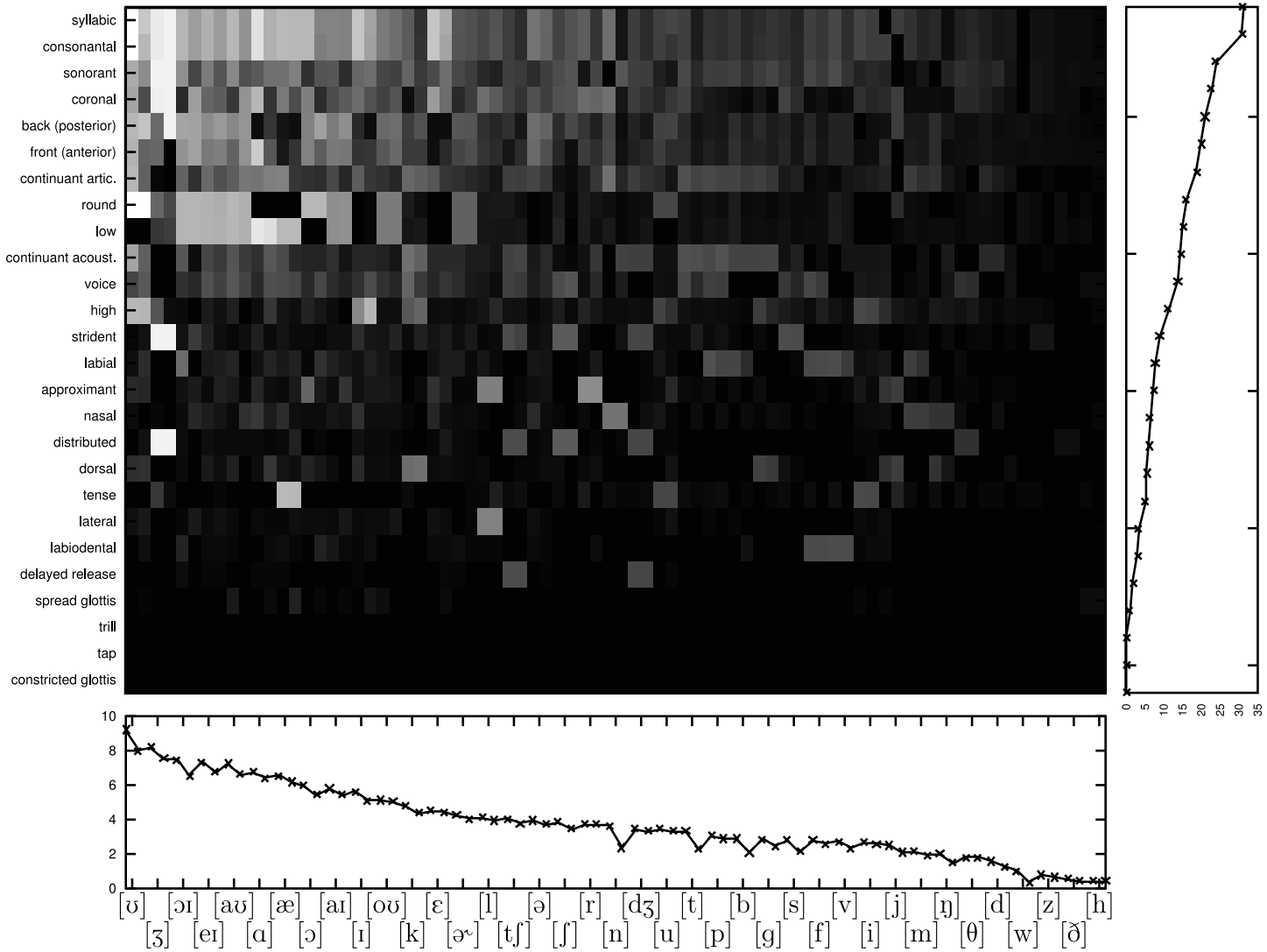


Figure 7.6: Variability of distinctive features within triphones for a given middle phone taken as reference. The distinctive features are sorted by their total variability (across all triphones) and the reference phones are sorted by the total variability across of its features within its set of triphones. The shades of gray represent the variability of a given feature within a give triphone. The white color correspond to the greatest variability and the black color to the smallest variability.

7.3 Length of Words

The study of language, under a phonological point of view, presupposes the discretization of the speech stream into unities, which are subject to rules. Phonological theories are based on the discretization of speech into segments that assembles themselves according to certain rules to create higher order structures. Words are assumed to be discrete entities built over other discrete units, such as syllables, morphemes and phonemes.

Words are argued to be unities of mental representation, although the very concept of word is still, in some aspects, unclear. Bloomfield (1926) introduced the concept of words as the smallest free form that may be uttered in isolation carrying a semantic or pragmatic content of its own, but it is doubtful since some words are not minimal free forms. It is fair to question whether or not the definite and indefinite articles, *the* and *a*, are themselves words at all. Do they carry a meaning of their own? As we have seen in previous examples, there are compound words, which carry just one meaning and requires more than one string of letters to express it. The compound word *blackboard* is just a combination of the words *black* and *board* that happens to designate an object that is not black at all. What amount the uncountable examples that an analytical language as German is capable of creating? The word for *blackboard* in German is *Tafel*, but it could also be *Wandtafel*, which is just a compound made of *Tafel* and *Wand* (Wall). The German word *Streichholzschachtel* (matchbook), is made of *streichen* (scratch), *Holz* (wood), *Schachtel* (box), while the English word *matchbox* is also a compound but made of *match* (not the Football one) and *box*. The way they are built is quite different, but make reference to the same object. That is something that tells about the way languages are structured and used, and it might be possible that the structures are a bit different in every language.

Olson (1994) stresses the point that the concept of the word as distinct unit is a by-product of literacy acquisition. The analysis of antique texts revels that texts were commonly redacted in *scriptura continua*, that means 'without word boundaries' (Saenger, 1991). The continuous string of characters was just a rightful representation of what speech utterances truly are. According to Saenger (1997), the lack of boundaries between words causes two drawbacks in the reading process: it slows down reading and encourages vocal activity. If the writing is dense and devoid of 'distinguishing signs', the decoding process is considerable longer and rely on vocalization. Silent and rapid reading is an achievement of our civilization which resulted from the development of a system of graphic signs in which the blank space is of prime importance. "In the sixth century, all manuscripts were still copied in *scriptura continua*, and it was not until between the sixth and eighth centuries that the separation of words was progressively introduced into all Latin manuscripts" (Pombo, 2002). Literacy is indeed an influential ability in our

language skills. As pointed by Hagège (1986) "writing is a linguistic analysis in various degrees of awareness".

There are many interactions between speech and writing. Lev S. Vygotsky was a psychologist who took an active interest in the cognitive consequences of writing, studying how speech affected writing and vice versa. "Writing requires deliberate analytic action on the part of the speaker[1]. In speaking, he is hardly conscious of the sounds he produces and quite unconscious of the mental operations he performs. In writing, he must take cognizance of the sound structure of each word, dissect it, and reproduce it in alphabetic symbols, which he must have studied and memorized before"(Vygotsky, 1934). The relationship between writing systems and spoken language is also a theme covered by Coulmas (2003). According to him, "the introduction of writing implies a cognitive reorientation and a restructuring of symbolic behaviour. Names of objects are conceptually dissociated from their denotata, as signs of physical objects are reinterpreted as signs of linguistic objects, names. In a second step, signs of names are recognized as potentially meaningless signs of bits of sound, which are then broken down into smaller components" (Coulmas, 2003).

Considering words as unities of mental processing, it is important to investigate the aspects involving this hypothesis. Miller (1956) suggested that the short-term memory storage capacity is constant in terms of the number of chunks. If we could consider words as chunks, then the short-term memory capacity should be the same regarding the size or duration of words. Baddeley et al. (1975) explores the relations between the memory span[2] and length of words. They observed that memory span is inversely proportional to word's length. Word's duration was recognized as an important aspect, since it was recognized that words of short temporal duration were better recalled than words of long duration, even when the number of syllables and phonemes are held constant. The results achieved by Baddeley et al. (1975) have some implications on Miller (1956)'s suggestions, "that memory span is limited in terms of number of chunks of information, rather that their duration. It suggests a limit to the generality of the phenomenon which Miller discusses, but does not, of course, completely negate it. The question remains as to how much of the data subsumed under Miller's original generalization can be accounted for in terms of temporal rather than structural limitations"(Baddeley et al., 1975).

Neath and Nairne (1995) points that "current explanations of the word-length effect rely on a time-based decay process within the articulatory loop structure in working

[1]Originally Lev S. Vygotsky was talking about *child*, but we made the substitution to comprise a wider context which we believe is still valid.

[2]Memory span is a common measure for short-term memory, where it is related to the length of a list of discrete items a person is capable of memorize and repeat back in order after presentation with an accuracy equals of superior to 50% of all trials. It appears to measure one's capacity to successfully distribute his attention and organize the incoming stimuli as a working unit.

memory", what does not completely explain one of the observations made by (Baddeley et al., 1975): "when articulation is suppressed by requiring the subject to articulate an irrelevant sound, the word length effect disappears with visual presentation, but remains when presentation is auditory". Neath and Nairne (1995) concludes that "word-length effects do not offer sufficient justification for including time-based decay components in theories of memory". He proposes then a feature model (Nairne, 1988, 1990) where the interferences handles the explanation of the observed data. It was designed to account for the major effects observed in immediate memory settings, what includes the recency effect, the effects of articulatory suppression, temporal grouping, and phonological similarity, among others.

In order to better understand the role played by these aspects into the way a language is structured, organized and used, we propose here a statistical analysis using a corpus. It would be time-spending and would require a great amount of work to collect a speech corpus and make use of it. Instead, we propose the usage of a text corpus, pronunciation dictionary and speech samples provided by online dictionaries. The analysis here will concern only the statistical aspects of written and spoken words length, what is important as length is regarded as an aspect of mental representation, among other features (Port, 2007).

Mendenhall realized that the study of word length, specifically, the analysis of the distribution of words of different lengths was important to establish comparisons of styles. Mendenhall (1887) investigated the differences in the literary styles of Dickens and Thackeray insofar as word-length distribution was concerned. The same approach was afterwards used (Mendenhall, 1901) to analyze the authorship of Shakespeare's plays[3]. In every Shakespeare's play the count of words of length four was always greater than the count of words of length three. Comparing with Bacon, the count of words of length three was greater than four, and Bacon also present a distinctly higher proportion of longer words than Shakespeare.

Apostel et al. (1957) supposes that words are built of a sequence of elementary unities and the process of assembling those unities into a speech stream is an elementary additive process. The cost assigned to this process is a compound of the number of building blocks used and their characteristics. The probability of occurrence of each word is assigned so that, on a long sequence, the average word-information-content is maximized, being also subject to the additive costs of building each unity and the sum of all the relative frequency of occurrence (probabilities) of each word must be one. Making the choice

[3]The Shakespeare authorship question was first posed in the middle of the 19th century, when the flattery of Shakespeare as the greatest writer of all time had become widespread. More than 70 authorship candidates have been proposed, including Francis Bacon.

that leads to such maximization is performing an entropy[4] maximization. The sender should maximize the transmission rate of information and concomitantly minimize the cost of transmission. This cost was expressed as a function of the length required in the transmission process, making short words a preferred choice. Mandelbrot demonstrated that a word-by-word encoding of a message will lead to the observed rank-frequency relation in natural languages.

Using our database we are able to estimate the number of words for a given length and also how frequently words of this length are used. Figure 7.7 presents these results. When the length of a word increases, the number of possible combinations of building unities increases factorially, but what we observe is an increase as short length words become longer and a great decrease after a turning point. The usage of shorter words, as expected, is greater, and the small drop observed for small words is just a consequence of the fewer number of possible combination leading to the existing words with short length words. Taking the average word length across words rank, as pictured in Figure 6.14c and 6.14d, the words with smaller rank are on average also smaller, either in the number of letters, phones or syllables.

Figures 7.9 and 7.10 present how the usage of words and phones is different when analyzing words of different length (number of letters, number of phones or number of syllables). It is possible to verify that in the extreme cases, when we analyze words of short length or long length, the pattern observed in the rank vs. frequency plot is quite different from the middle length words. The short and long words experience steeper decay regarding the occurrence of words or phones.

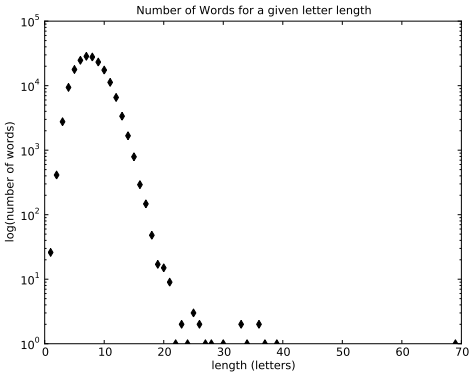
The relationship between the number of phones and the number of letters and the number of phones and the number syllables may be observed in Figure 7.11. Phones and Letters have a straight line relationship, and the majority of the words have between 3 and 9 letters and 3 and 10 phones. The relationship between number of phones and syllables is a little more dispersed, and we may observe that the majority of words have between 1 and 3 syllables, corresponding to the range of 3 to 9 phones.

When considering the frequency of usage of words with a certain number of syllables and phones, or phones and letters, we observe a more diffuse relation that still resembles the relations regarding the number of words. This might be observed in Figure 7.12.

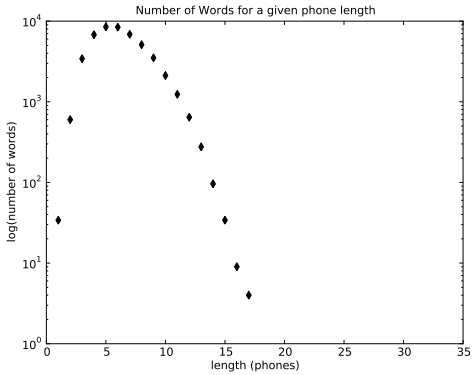
Another way to regard words length is to measure its duration when uttered. Although each speaker will utter the same word with a different duration at each trial, on the average, and under normal conditions, we expect to observe the same duration, what

[4]As defined by Shannon, the entropy is an additive measure of the amount of available choice in the process of selecting one from the allowable messages among many in each unit of the information transmission process. Entropy is used as a measure of the amount of information carried by a random variable.

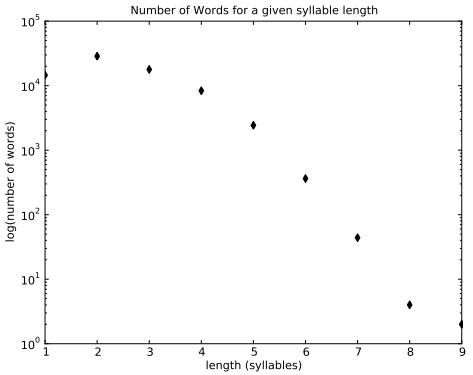
might not differ much from one speaker to another, unless one suffer from some sort of aphasia, compromising the speaker elocution capabilities, or he is under a certain situation in which he is required to speak on a much slower or faster rate. The data then here analyzed consists of utterance duration of words collected from online dictionaries: Cambridge, Dictionary.com, The free dictionary and Macmillan. The samples where collected from each dictionary and its duration were afterwards normalized within each group. Unfortunately each dictionary has samples from different speakers and there is no control under the speech rate used. Although the data don't present a strong correlation it is not so weak and we might observe, in Figure 7.13, that the relation between number of syllable and duration of the uttered word is more diffuse than the previous relations presented before, but still there is a straight line tendency.



(a) number of letters.

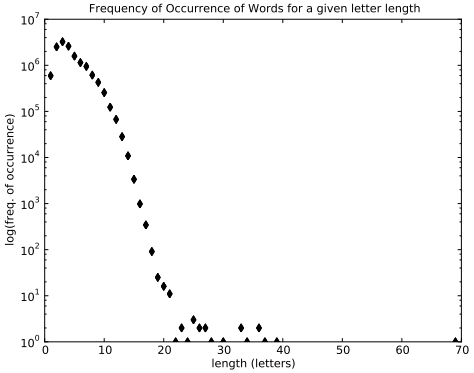


(b) number of phones

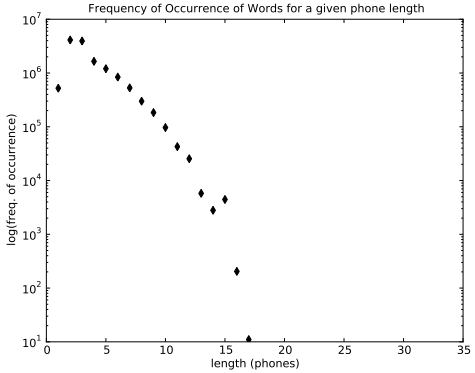


(c) number of syllables

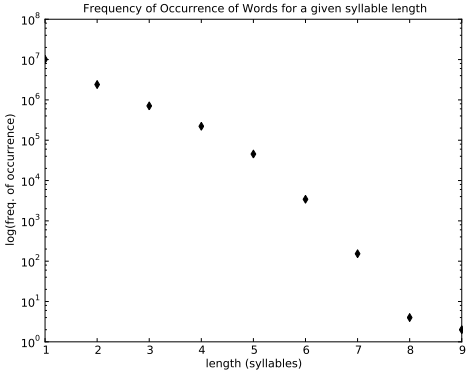
Figure 7.7: Number of words in English with a certain length, measured as the number of letters, phones or syllables.



(a) number of letters.

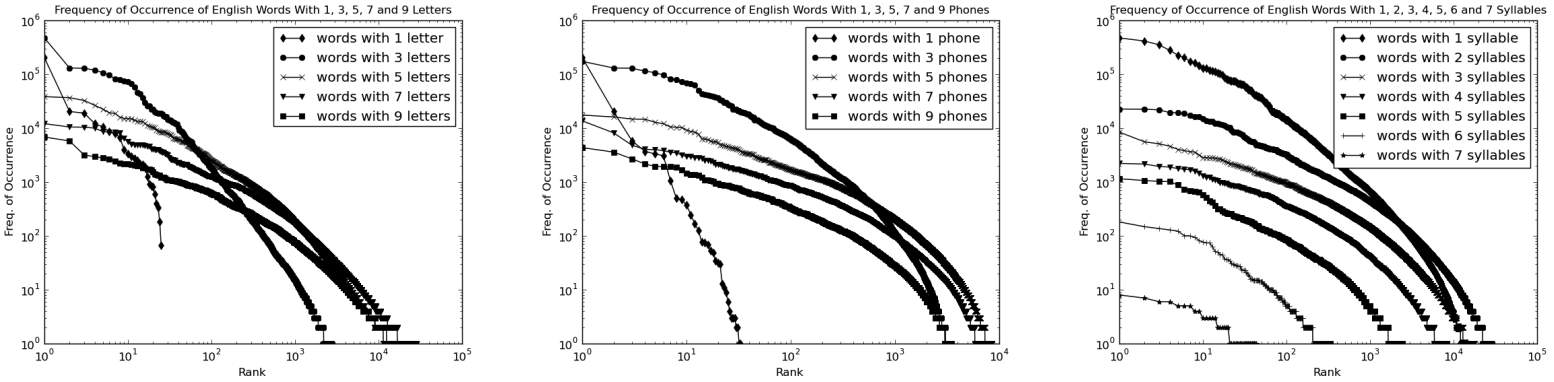


(b) number of phones



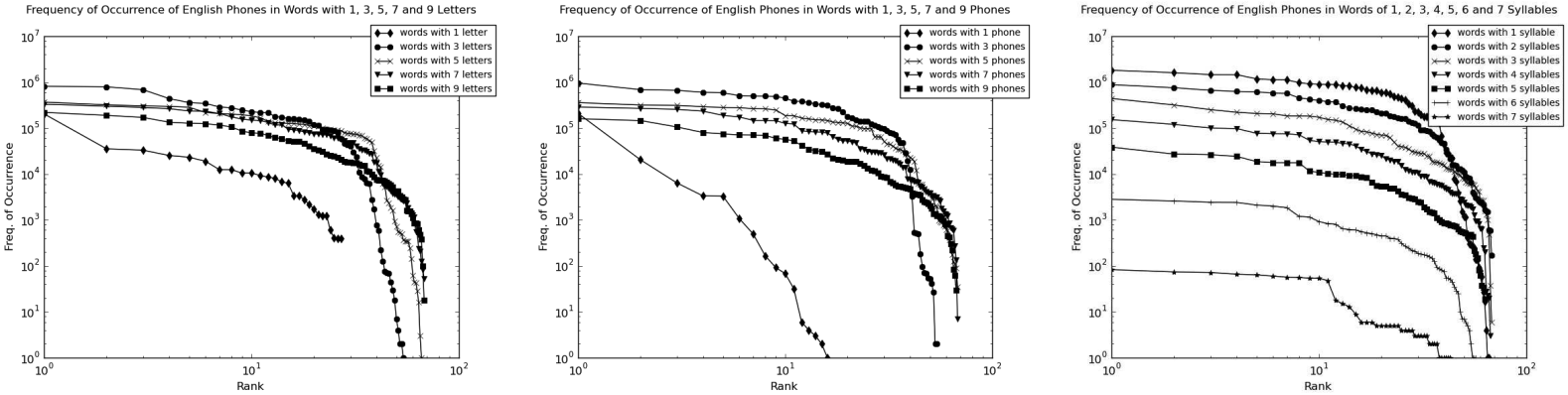
(c) number of syllables

Figure 7.8: Frequency of occurrence of words in English with a certain length, measured as the number of letters, phones or syllables.



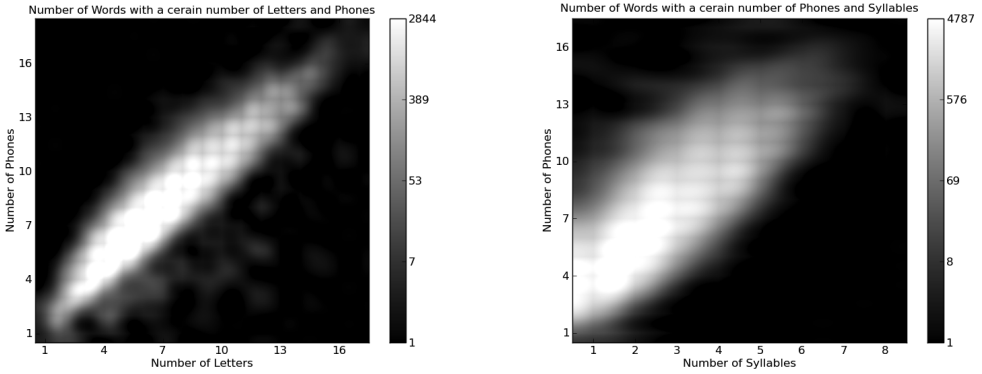
(a) number of letters. (b) number of phones (c) number of syllables

Figure 7.9: Frequency of occurrence of words in English with length in a certain range, measured as the number of letters(7.9a), phones(7.9b) or syllables(7.9c).



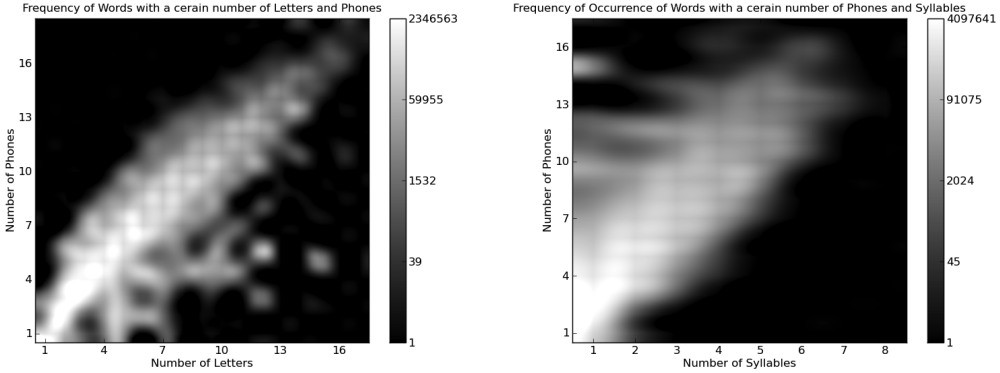
(a) number of letters. (b) number of phones (c) number of syllables

Figure 7.10: Frequency of occurrence of phones in English within words with length in a certain range, measured as the number of letters(7.10a), phones(7.10b) or syllables(7.10c).



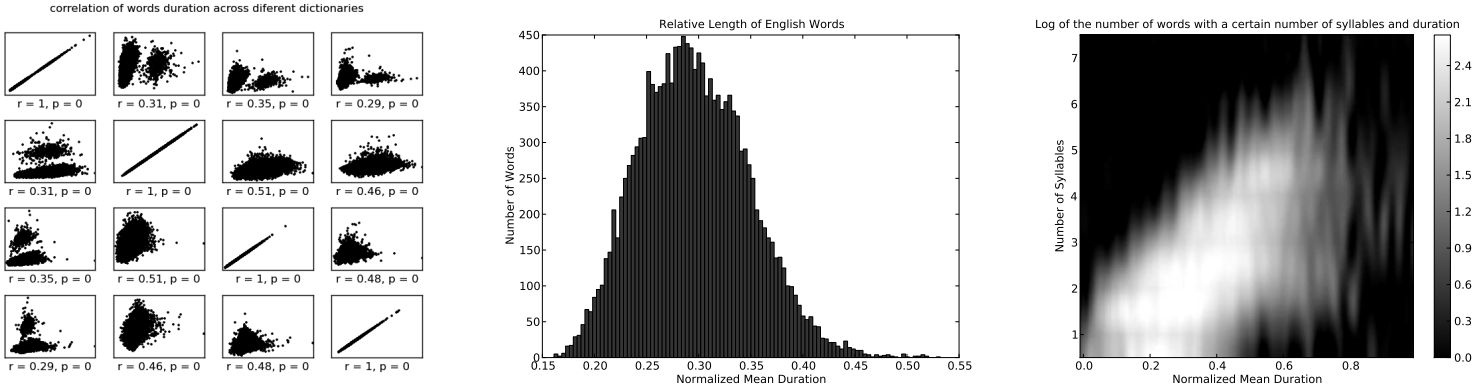
(a) Letters and Phones. (b) Phones and Syllables.

Figure 7.11: Number of English words with a certain number of letters, phones and syllables.



(a) Letters and Phones. (b) Phones and Syllables.

Figure 7.12: Frequency of occurrence of words in English with a certain number of letters, phones and syllables.



(a) Scatter plot of words utterance duration from four different dictionaries.

(b) Histogram of words mean (across the four different dictionaries) utterance duration.

(c) Number of words with a certain number of syllables and a certain duration.

Figure 7.13: The data here presented uses the utterances duration of speech samples from the four on-line dictionaries. The durations were first normalized within each dictionary. Only the words found in all the four dictionaries were used.

7.4 Menzerath's law

Menzerath (1928) observed the relation between the average syllable length and the word length, regarding the number of syllables that build the word, was a decreasing relationship, as the number of syllables in a word grew, the average length of the syllables decreased. In general form, this observation would further be formulated as a linguistic law according to which the increase of a linguistic construct results in a decrease of its constituents, and vice-versa. Examining the structures of German words, Menzerath (1954) stated the hypothesis: "Je größer das Ganze, desto kleiner die Teile" (the longer the whole, the smaller the parts). This hypothesis did not apply solely to Linguistics, but we are here concerned with its implications on the study of languages. That could imply that there is no uniformity in the units of a language and the hypothesis could be specified as "the more complex a linguistics' construct, the simpler are its constituents". This hypothesis seems plausible: linguistics's constructs carry information and are made of specific constituents. The constituents are chosen so that the construct may be clearly identified and they should have enough redundancy so that, even under severe interference, they might be distinctive one from another. For the reasons given, short construct should be made of longer and more distinctive constituents, while for longer construct shorter and less distinctive constituents might also be used, and the existence of redundancy should be adequately secured to properly differentiate one construct from another.

In order to derive the relation between constructs and constituents, we may investigate the length relationship between them. We may denote by y the length of the constituent and by x the length of the construct. The relative change rate on the length of the constituent, the first derivative of y divided by y, is inversely proportional to the length of the construct. The increase of the length of the constituent is also proportional to the length of the constituent. Mathematically, $y'/y \sim 1/x$, and we may then state in the following form:

$$\frac{y'}{y} = \frac{b}{x}. \tag{7.1}$$

This differential equation my be solved directly by integration, resulting in

$$\ln y = b \ln x + c, \tag{7.2}$$

From what follows that

$$y = ax^b \tag{7.3}$$

where $a = e^c$, and therefore, this parameter is always greater than zero, following then that the curve stated by Relation 7.3 is a rising convex curve when $b > 1$, it will be a concave rising curve when $0 < b < 1$ and it will be a convex falling curve when $b < 0$.

Altmann (1980) formulated mathematically the Menzerath's principle. He considered also a disturbance, making the relation become

$$\frac{y'}{y} = \frac{b}{x} + c, \tag{7.4}$$

from which the solution is stated as:

$$y = ax^b e^{cx} \ . \tag{7.5}$$

The curve will be monotonically decreasing when $-b/c > x$. Considering the solution of Relation 7.4, we might argue whether the perturbation arises or not from the parameters b or c. Taking $b = 0$, it will lead us to the third solution, which is

$$y = ae^{cx}. \tag{7.6}$$

If the observed pattern does not agree with the previous stated relations, it doesn't necessarily implies a falsification of the Menzerath's law, but it could imply that other factors should also be taken into consideration, for example, the position of a constituent in the construct Altmann et al. (1989). Let us denote by z the position of the constituent, meaning that it is the z-th constituent inside a construct. The differential equation could then be rewritten as

$$\frac{y'_x + y'_z}{y} = -\frac{b}{z} + c, \tag{7.7}$$

and the solution would be

$$y = az^b e^{cx}. \tag{7.8}$$

The more factors are taken into account, the more complete and complex will get our model.

We present bellow, in Figure 7.14, the relation between construct (word) length (considering the number of syllables or phones it is made of), and the average duration of its constituents (syllables or phones) or the average number of phones per syllables. The Menzerath model was fitted using a least-mean squares procedure and the parameter obtained are presented bellow or in the Figure 7.14, with the continuous fitted curve also displayed.

For the relation between words syllable length and average syllable duration the parameter found were: $a = 743$, $b = -0.916$ and $c = 0.072$. For this model, the Pearson's correlation coefficient found (between the data and the model's prediction) is 0.93. For the relation between words phone length and the average duration of phones, the parameter were: $a = 619$, $b = -0.901$ and $c = 0.039$. In this case, the correlation coefficient

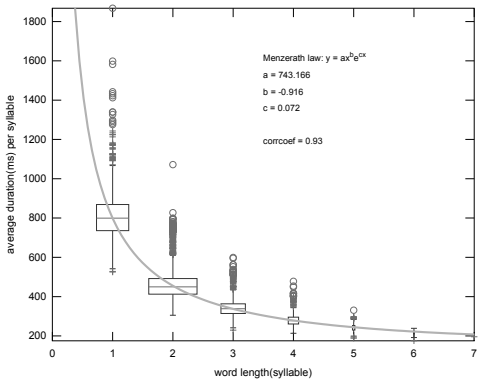
found was 0.91. For the last one, the relation between word syllable length and the average number of phones per syllable, the parameter found were $a = 3.22$, $b = -0.451$ and $c = 0.067$. The correlation coefficient was 0.49. All models show the parameter c close to zero, meaning that the simpler model, depicted in Equation 7.3, would be suitable. For the last case scenario, the correlation between the model and the data was poor, what leads that the model as proposed did not represent well the intrinsic relation, and maybe some further considerations should be taken, as previously discussed.

The Menzerath's relationship was previously presented in various levels of language units. Going towards an upper level of syntactic generalization, we could expect to see that same relationship on the word-sentence level. As depicted in Figure 7.15, it appears that an intermediate unit must be introduced between words and sentences (Köhler et al., 2005). This intermediate unit might be phrases or clauses, which are regarded as direct constituent of sentences, as stated by Grzybek et al. (2006). Figure 7.15 presents the relation between words length (average number of characters that words are made of in a given sentence) and sentence length (number of words in the given sentence). Each dot in this Figure represents a sentence in *Ulysses*, by James Joyce. We might observe that, as sentences get longer, the average word length tends to converge to the value of 4.46. That could be reasoned as the result of two interacting forces, one tending to make words shorter as the sentences grow longer, and another tending to make words longer. The equilibrium would be found in the average word length of 4.46.

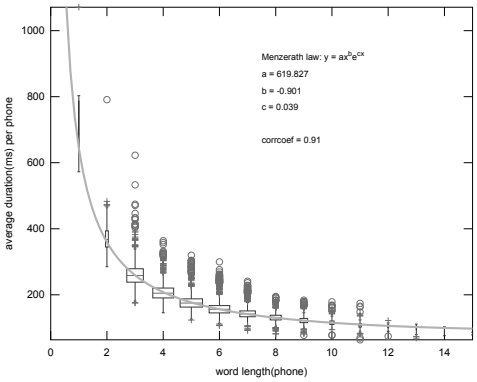
Altmann and Arens (1983), interpreting the results related to the Menzerath-Altman Law, pointed out that the relationship described by the law is more likely to hold true only if we are dealing with direct constituents of a given construct. From that observation and from the results observed in Figure 7.15, we might conclude that at least one intermediate unit in the linguistic construct levels is necessary. Buk and Rovenchak (2007) suggest the usage of clauses and present results from the analysis of Ukrainian texts. As a result, words may be considered as a direct constituent of clauses, but not of sentences.

The convergence of the average word length presented in Figure 7.15 might be a result of the low frequency presented in the analyzed text for long senteces, as might be observed in the frequency vs. sentence length plot in Figure 7.16. It might also be that what drives this convergence has a psycholinguistic motivation rather than an statistical one. Regarding the human processing limits, as presented by Miller (1956), there is a magical rule of 7 ± 2, which could serve as a limitation on the average length of words in phrases or clauses.

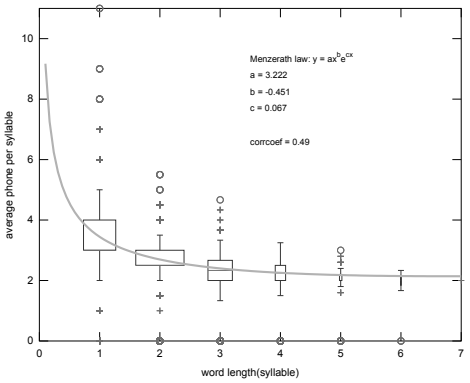
Investigating the relation between sentences length and average words length, but this time, using the number of syllables as the unit to measure a word length, we observe quite a different behavior, as it is presented in Figure 7.17. The analysis was again based on



(a) Average duration of syllables as the number of syllables in a word increases.



(b) Average duration of phones as the number of phones in a word increases.



(c) Average number of phones per syllables as the number of syllables in words increases.

Figure 7.14: The pronunciation duration of each word is the average of the speech sample from four different online dictionaries (Cambridge Dictionary, Dictionary Reference, The Free Dictionary and Macmillan Dictionary). The Boxplots display the relation between the words length (number of phones or number of syllables) and the average duration of its constituent parts (phones or syllables). The last Figure also presents the relation between the average number of phones per syllables as a function of the syllable length of words.

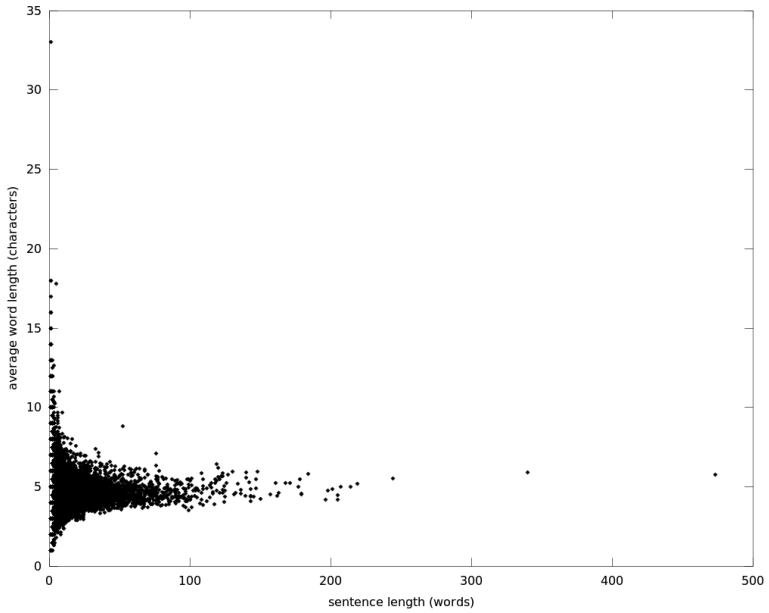


Figure 7.15: Average word length (number of letters) versus sentence length (number of words). The mean word length average value is 4.46.

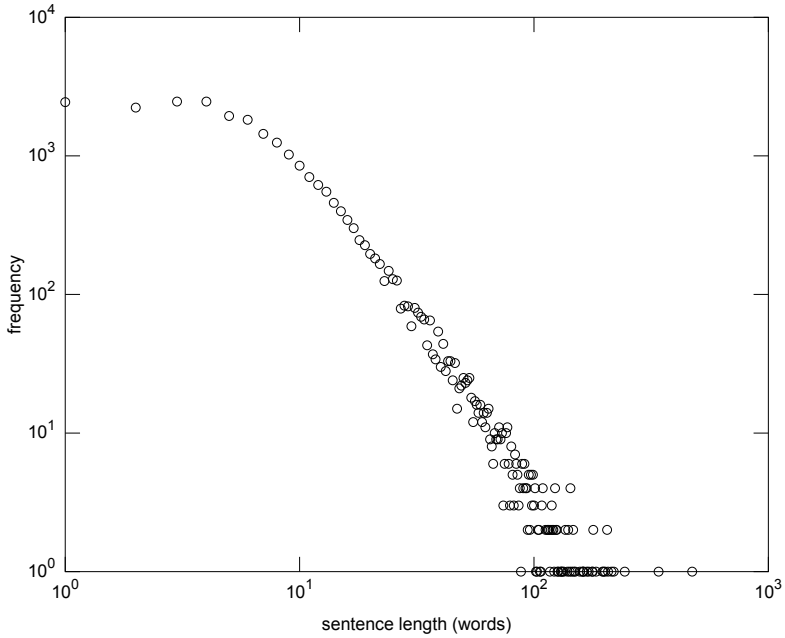


Figure 7.16: Frequency of occurrence of sentences for a given length.

the text *Ulysses*. Each word, in each sentence, was transcribed into syllables using online dictionaries, and we considered only the sentences where every word could be transcribed. The number of words and syllables in each sentence was drawn, and the result is presented in Figure 7.17. The pattern presented in that figure suggests a quantization effect of the

size of half a syllable. This might indicate that demi-syllables are actual constituent units of language, what would corroborate the good achieved results in speech recognition systems using demi-syllables as the base unit (Yoshida et al., 1989).

In speech recognition it is important to have a unit smaller than a word (Shoup, 1980). The demi-syllable is defined as a half syllable, divided at the center of the syllable nucleus. This unit holds a transitional information, which is important in the speech recognition task. It implicitly holds phonological constrictions, it is also suitable in size and can take into account the co-articulatory effects.

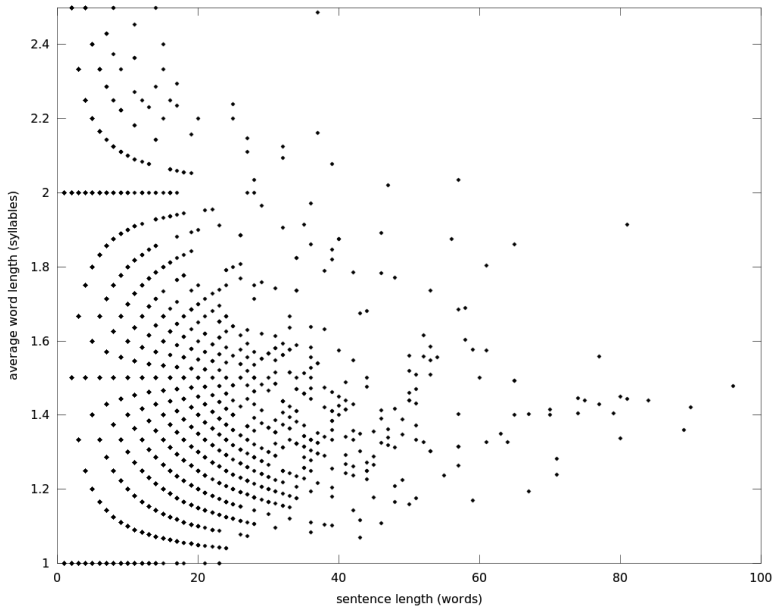


Figure 7.17: Relation between sentences length (number of words) and the average word length (number of syllables a word is made of).

7.5 Zipf Fit

In the previous Chapter, we saw the Zipf plot of words, phones, diphones, triphones and syllables. One way to compare the Zipf law for different data sources is to find the Zipf exponent that best fits the given data. Once we have the model established and the data that we believe might be explained by such model, we need a procedure to find the parameters that adjusts the best model for the given data. There are two general methods for parameter estimation: least-squares estimation (LSE) and maximum likelihood estimation (MLE). The former is very popular and is tied to some familiar statistical concepts, such as linear regression. LSE requires no distributional assumption on the data and is very helpful for obtaining a descriptive model summarizing the observed

data, but "it has no basis for testing hypotheses or constructing confidence intervals. (...) MLE has many optimal properties in estimation: sufficiency (complete information about the parameter of interest contained in its MLE estimator); consistency (true parameter value that generated the data recovered asymptotically, i.e. for data of sufficiently large samples); efficiency (lowest-possible variance of parameter estimates achieved asymptotically); and parameterization invariance (same MLE solution obtained independent of the parametrization used). In contrast, no such things can be said about LSE."(Myung, 2003).

The Zipf model states that the probability of occurrence of samples is determined by a power law relation and the rank of each sample. Having defined the model, we want to find the population, defined by a corresponding distribution, from which it is most likely that our observed samples would come from. Associated with each population is a unique parameter for the model assumed. The Zipf model state that the probability density function (PDF) is defined by

$$f(k|s,N) = \frac{1/k^s}{\sum_{n=1}^{N} n^{-s}} = \frac{1/k^s}{H_{N,s}}, \tag{7.9}$$

where k is the rank of the observed samples, s is the model parameter we want to determine, and N is the number of observations. If individual observations are statistically independent one from another, then the PDF of the set of observed data is equal to the product of all individual observations PDFs. Suppose we have M observations, then we will have the following probability of experiencing this set of samples

$$\begin{aligned} f(k=(k_1,k_2,\cdots,k_M)|s,N) &= \prod_{m=1}^{M} f(k_m|s,N) \\ &= \prod_{m=1}^{M} \frac{1}{k_m^s} \frac{1}{H_{N,s}} \\ &= \left(\frac{1}{H_{N,s}}\right)^M \prod_{m=1}^{M} \frac{1}{k_m^s}. \end{aligned} \tag{7.10}$$

Given a model and its parameters value, we might show that some data are more probable than others, we appraise this through the PDF. As we are faced with the data and we want to find the model that is the most probable to have generated that data, we are dealing with the inverse problem. A *likelihood function* is then defined by reversing the roles of the data and the parameters in $f(k|s,N)$, i.e.

$$L(s|k,N) = f(k|s,N). \tag{7.11}$$

It represents the likelihood of the parameter s given the observed data, and as such it is a function of s.

The Figure 7.18 shows the likelihood function $L(s|k_m, N)$ for different values of k, the rank. As there are many observations, we conclude that the likelihood function which takes into account all observations is the product of all the likelihood functions

$$\begin{aligned} L(s|k_1, \cdots, k_M, N) &= \prod_{m=1}^{M} L(s|k_m, N) \\ &= \left(\frac{1}{H_{N,s}}\right)^M \prod_{m=1}^{M} \frac{1}{k_m^s}. \end{aligned} \tag{7.12}$$

From a simple visual observation of Figure 7.18 we may guess that the parameter s that gives the maximal likelihood for a set of observations might be somewhere in the interval $[1, 2]$.

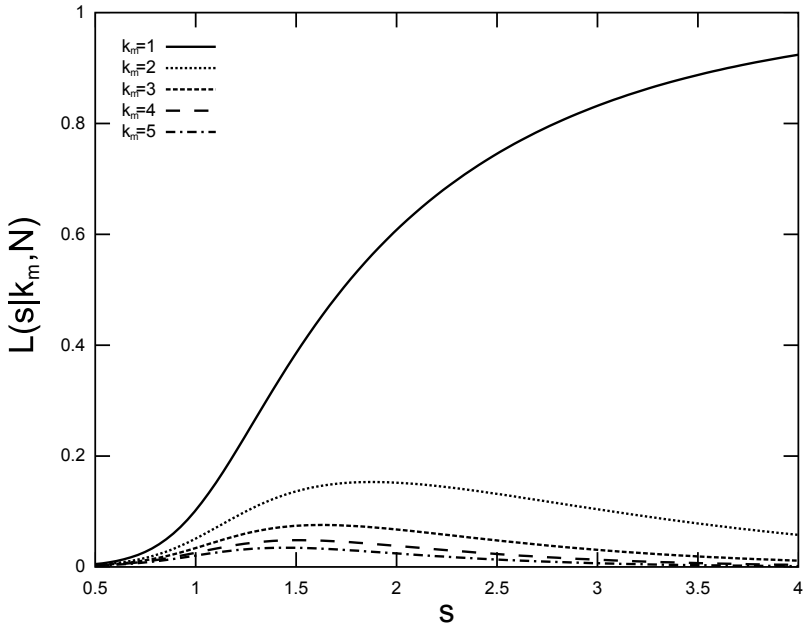


Figure 7.18: Likelihood terms for the Zipf model ($N = 100$).

The principle of maximum likelihood estimation (MLE) was introduced by Fisher (1922), where he states that the desired probability distribution is the one that makes the observed data most likely. The parameter that maximizes this likelihood function is the MLE estimate (Aldrich, 1997).

The MLE estimate might not exist or even may not be unique. For computational convenience, the MLE estimate s_{MLE} is obtained by maximizing the log-likelihood function.

Assuming that $\ln L(s|k,N)$ is differentiable, the maximal value takes place when

$$\frac{\partial \ln L(s|k,N)}{\partial s} = 0 \tag{7.13}$$

and when the second derivative is negative

$$\frac{\partial^2 \ln L(s|k,N)}{\partial s^2} < 0. \tag{7.14}$$

The logarithm of the likelihood function is given by

$$\ln L(s|k_1,\cdots,k_M,N) = -M \ln H_{N,s} - s\sum_{m=1}^{M} \ln k_m. \tag{7.15}$$

We need then to solve the equation 7.13. So we have to solve

$$\begin{aligned}
\frac{\partial \ln L(s|k_1,\cdots,k_M,N)}{\partial s} &= -M\frac{d}{ds}\ln H_{N,s} - \sum_{m=1}^{M}\ln k_m \\
&= -M\frac{1}{H_{N,s}}\frac{d}{ds}H_{N,s} - \sum_{m=1}^{M}\ln k_m \\
&= -M\frac{-\sum_{n=1}^{N} n^{-s}\ln n}{\sum_{n=1}^{N} n^{-s}} - \sum_{m=1}^{M}\ln k_m \\
&= -M\frac{G_{N,s}}{H_{N,s}} - \sum_{m=1}^{M}\ln k_m = 0
\end{aligned} \tag{7.16}$$

We need to find s that satisfies the equation 7.16 and for which the second derivative of the log-likelihood is negative. Taking the second derivative of the log-likelihood, we have

$$\begin{aligned}
\frac{d^2}{ds^2}\ln L(s|k_1,\cdots,k_M,N) &= \frac{d}{ds}\left(-M\frac{G_{N,s}}{H_{N,s}}\right) \\
&= -M\frac{\frac{dG_{N,s}}{ds}H_{N,s} - G_{N,s}\frac{dH_{N,s}}{ds}}{H_{N,s}^2} \\
&= -M\frac{\left(\sum_{n=1}^{N} n^{-s}\ln^2 n\right)H_{N,s} - G_{N,s}^2}{H_{N,s}^2} \\
&= M\frac{G_{N,s}^2 - I_{N,s}H_{N,s}}{H_{N,s}^2}
\end{aligned} \tag{7.17}$$

The denominator in Equation 7.17 is always positive, so we need to verify if $I_{N,s}H_{N,s} \geq G_{N,s}^2$, in order to make the second derivative negative.

We might then write

$$\begin{aligned} G_{N,s}^2 &= \left(\sum_{n=1}^{N} n^{-s}\ln n\right)^2 = \left(\sum_{n=1}^{N} n^{-s/2}(n^{-s/2}\ln n)\right)^2 \\ &\leq \left(\sum_{n=1}^{N}\left(n^{-s/2}\right)^2\right)\left(\sum_{n=1}^{N}\left(n^{-s/2}\ln n\right)^2\right) \\ &= \left(\sum_{n=1}^{N} n^{-s}\right)\left(\sum_{n=1}^{N} n^{-s}\ln^2 n\right) \\ &= H_{N,s}I_{N,s}\ , \end{aligned} \tag{7.18}$$

where we have used the Cauchy-Schwarz inequality.

It has been proved that $I_{N,s}H_{N,s} \geq G_{N,s}^2$ and therefore the second derivative of the likelihood function is always negative. Any s that satisfies equation 7.16 is a maximum likelihood estimation. We may use any root finding algorithm to find the MLE parameter of the Zipf model.

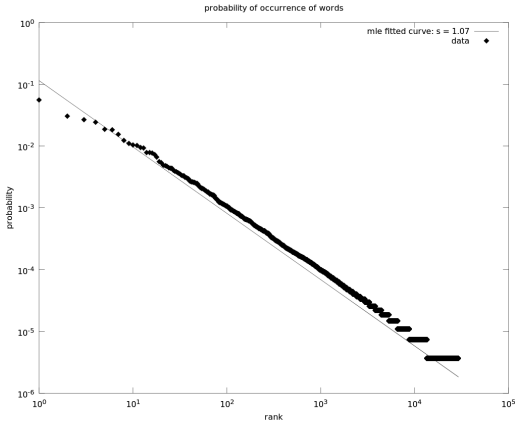
Figure 7.19 presents some examples where we might observe that the value of s_{MLE} for a natural text is close and above 1. All other random synthetic texts present a s_{MLE} whose value is bellow 1. As the source characteristics approach a white random source, the closer the estimated parameter gets to zero. A similar behavior might be observed when analyzing the entropy of the source. We shall see that the entropy increases as the characteristics of the source approach a white random process. Another aspect about the distinction on the model coefficient is important to note. The value of 1 seems to be a water shed between natural and random processes. From the Zipf relation in equation 7.9 we might observe that for large extremely vocabularies the value of the exponent must be above 1 so that the normalizing coefficient $H_{N,s}$ converge. $H_{N,s}$ is a Riemann zeta function which converges when the real part of the exponent s is greater than 1.

7.5.1 Zipf-Mandelbrot Fit

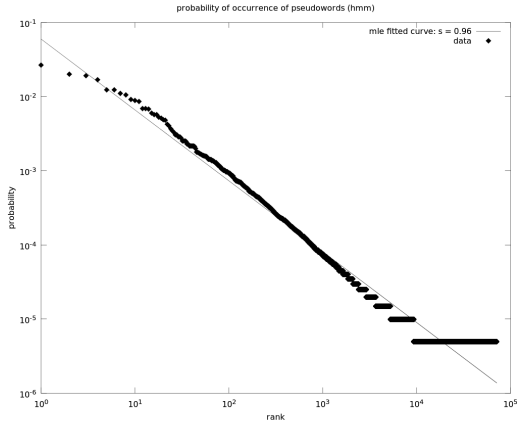
The same sort of procedure might be used to fit a Zipf-Mandelbrot model, given a set of observed values. The Zipf-Mandelbrot distribution is given by

$$f(k|s,q,N) = \frac{1/(k+q)^s}{\sum_{n=1}^{N}(n+q)^{-s}} = \frac{1/(k+q)^s}{H_{N,s,q}} \tag{7.19}$$

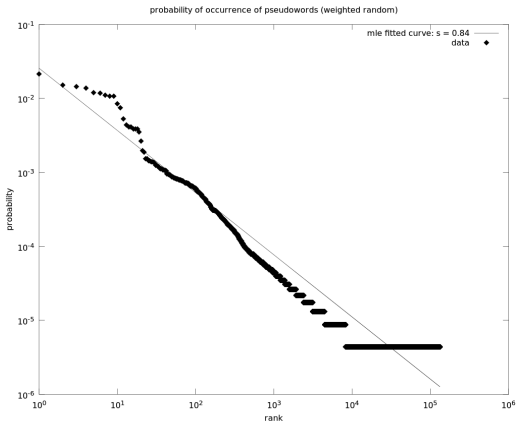
where q denotes the flattening constant and $H_{N,s,q}$ is similar to the generalized harmonic number.



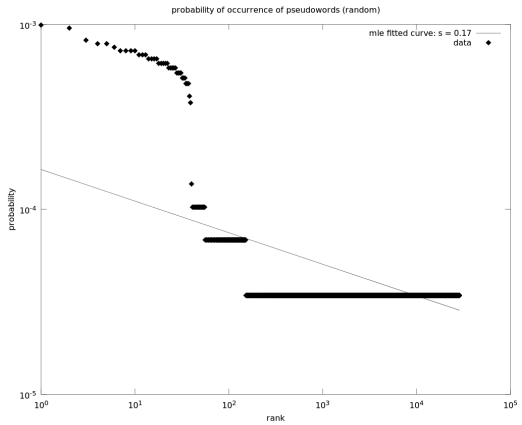
(a) Natural text. $s_{MLE} = 1.0738$



(b) Random text created by a Markov Model. $s_{MLE} = 0.95574$



(c) Random text with symbols weighted probabilities. $s_{MLE} = 0.84042$



(d) White random text. $s_{MLE} = 0.17076$

Figure 7.19: Zipf plot of the frequency of occurrence of (pseudo)words and the fitted model.

For a set of M observations, $k_1, k_2, \ldots, k_m$, the likelihood function will be given by

$$L(s|k_1, \cdots, k_M, N) = \left(\frac{1}{H_{N,s,q}}\right)^M \prod_{m=1}^{M} \frac{1}{(k_m+q)^s}. \tag{7.20}$$

and the logarithm of the likelihood is given by

$$\ln L(s|k_1, \cdots, k_M, N) = -M \ln H_{N,s,q} - s \sum_{m=1}^{M} \ln(k_m+q) \ . \tag{7.21}$$

In order to find the maximum likelihood estimates for s and q, we need to find

$$\frac{\partial \ln L(s|k_1, \cdots, k_M, N)}{\partial s} = 0 \qquad \frac{\partial^2 \ln L(s|k_1, \cdots, k_M, N)}{\partial s^2} < 0 \quad (7.22,\ 7.23)$$

and

$$\frac{\partial \ln L(s|k_1, \cdots, k_M, N)}{\partial q} = 0 \qquad \frac{\partial^2 \ln L(s|k_1, \cdots, k_M, N)}{\partial q^2} < 0 \ . \quad (7.24,\ 7.25)$$

Lets first consider the parameter s. The partial derivative of the log likelihood in relation to s is given

$$\begin{aligned}
\frac{\partial \ln L(s|k_1, \cdots, k_M, N)}{\partial s} &= -M \frac{1}{H_{N,s,q}} \frac{\partial H_{N,s,q}}{\partial s} - \sum_{m=1}^{M} \ln(k_m+q) \\
&= -M \frac{1}{H_{N,s,q}} \frac{\partial}{\partial s} \sum_{n=1}^{N} (n+q)^{-s} - \sum_{m=1}^{M} \ln(k_m+q) \\
&= -M \frac{1}{H_{N,s,q}} \left(-\sum_{n=1}^{N} (n+q)^{-s} \ln(n+q) \right) - \sum_{m=1}^{M} \ln(k_m+q) \\
&= -M \frac{G_{N,s,q}}{H_{N,s,q}} - \sum_{m=1}^{M} \ln(k_m+q)
\end{aligned} \tag{7.26}$$

and the second derivative is given by

$$\begin{aligned}
\frac{\partial^2 \ln L(s|k_1, \cdots, k_M, N)}{\partial s^2} &= -M \frac{\partial}{\partial s} \left(\frac{G_{N,s,q}}{H_{N,s,q}} \right) \\
&= -M \frac{\frac{\partial G_{N,s,q}}{\partial s} H_{N,s,q} - G_{N,s,q} \frac{\partial H_{N,s,q}}{\partial s}}{H_{N,s,q}^2} \\
&= -M \frac{\left(-\sum_{n=1}^{N} (n+q)^{-s} \ln^2(n+q) \right) H_{N,s,q} - G_{N,s,q}^2}{H_{N,s,q}^2} \\
&= M \frac{G_{N,s,q}^2 - I_{N,s,q} H_{N,s,q}}{H_{N,s,q}^2}
\end{aligned} \tag{7.27}$$

and it is shown as negative by following the same steps as in Equation 7.18.

Considering now the parameter q, we also need to calculate the first and second derivatives of the log likelihood in relation it. The first derivative is given by

$$\begin{aligned}\frac{\partial \ln L(s|k_1,\cdots,k_M,N)}{\partial q} &= -M\frac{1}{H_{N,s,q}}\frac{\partial}{\partial q}H_{N,s,q} - s\sum_{m=1}^{M}\frac{1}{k_m+q} \\ &= Ms\frac{1}{H_{N,s,q}}\sum_{n=1}^{N}(n+q)^{-s-1} - sH_{N,1,q} \\ &= sM\frac{H_{N,s+1,q}}{H_{N,s,q}} - sH_{N,1,q} \end{aligned} \tag{7.28}$$

and the second derivative is given by

$$\begin{aligned}\frac{\partial^2 \ln L(s|k_1,\cdots,k_M,N)}{\partial q^2} &= sM\frac{\frac{\partial H_{N,s+1,q}}{\partial q}H_{N,s,q} - H_{N,s+1,q}\frac{\partial H_{N,s,q}}{\partial q}}{H_{N,s,q}^2} - s\frac{\partial}{\partial q}\sum_{n=1}^{N}(n+q)^{-1} \\ &= sM\frac{\left(\sum_{n=1}^{N}\frac{-s-1}{(n+q)^{s+2}}\right)H_{N,s,q} - H_{N,s+1,q}\left(\sum_{n=1}^{N}\frac{-s}{(n+q)^{s+1}}\right)}{H_{N,s,q}^2} + \dots \\ & \quad s\sum_{n=1}^{N}(n+q)^{-2} \\ &= sM\frac{-(s+1)H_{N,s+2,q}H_{N,s,q} + sH_{N,s+1,q}^2}{H_{N,s,q}^2} + sH_{N,2,q} \\ &= s^2M\frac{H_{N,s+1,q}^2}{H_{N,s,q}^2} - s(s+1)M\frac{H_{N,s+2,q}}{H_{N,s,q}} + sH_{N,2,q}\ . \end{aligned} \tag{7.29}$$

The second derivate in relation to q is also negative[5].

We conclude that the parameters that satisfies equations 7.22 and 7.23 for s and equations 7.22 and 7.23 for q are the maximum likelihood estimates for the Zipf-Mandelbrot model given the observed data. They might be found by any root-finding algorithm.

7.6 Inverse Zipf

Zipf (1935) states the inverse law, which relates the number of words for a given frequency to the frequency of occurrence by a power law function as described by the following equation

$$N(f) = af^{-b} \tag{7.30}$$

[5]This was not formally proved, but computation with various values of s, q and N indicates that it is indeed negative.

where N stands for the number of words with a given frequency f. The constants a and b are parameters of the model. In the case of natural texts, the value of b is usually close to 2.

Imagine we have a large corpus where we might observe the occurrence of M different words W. The lexical described by this corpus is the set $\{W_1, W_2, \ldots, W_M\}$. In this corpus we verify that each word presents a different frequency of occurrence, that is described by f_W. The hypothetical frequency observed in a near-infinite corpus is called the "underlying frequency". The simplest way to estimate the frequency of occurrence is to count the number of occurrences and divide by the total counts. This estimator is in fact the maximum likelihood estimator. The probability of occurrence of a word is then given by

$$p_W = \frac{f_W}{N} \tag{7.31}$$

where f_W is the count of occurrence for a given word W and N is the total number of words in the corpus, $N = \sum_{m=1}^{M} f_{W_m}$. That would lead then to the satisfaction of the relation $\sum p_W = 1$. But this cannot be, since, no matter how large the corpus is, there will always be words with zero frequency in this corpus.

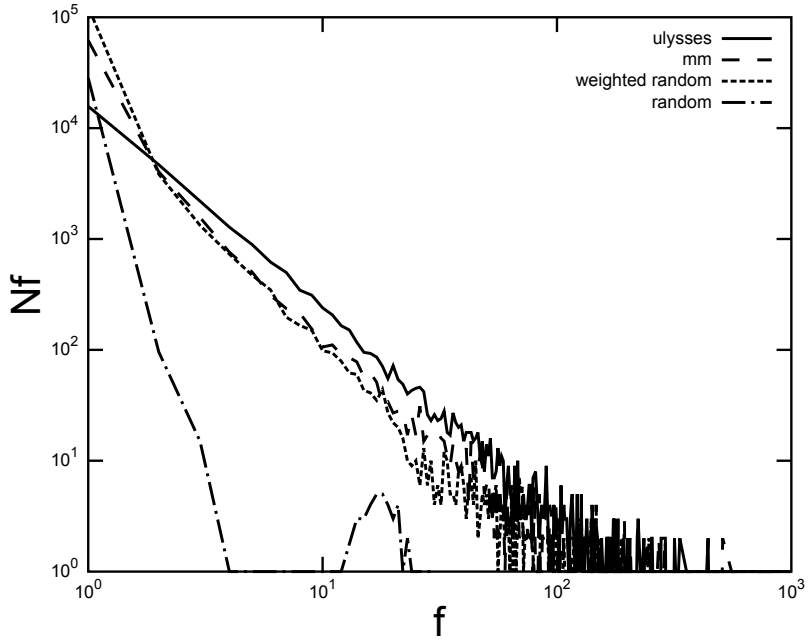


Figure 7.20: Inverse Zipf plot applied to Ulysses data (words) and random generated words. The random texts are created using the procedure as described before.

The Figure 7.20 presents the relation between the frequency of occurrence of words f and the number words which occur with that given frequency. We might observe that words that happen few times are in a great number, and there are just a few words which

are very frequent. At high frequencies some gaps will emerge, that means there is no word with that given frequency of occurrence. We might observe these gaps in the list of words and their frequency. The most frequent word *the* happens 15,126 times in *Ulysses*; the next most frequent word *of* happens 8,256. There will be no word filling the gap between 15,126 and 8,256. These gaps are also present in random generated text, but they are far less common. Analyzing the texts in the example of Figure 7.20, we may count the percentage of frequencies which present no occurring (pseudo-)words, the result is 50,8%, 7,1%, 3,5% and 0,0% for the natural text, Markov model, weighted random and random generated texts, respectively. The difference presented in the percentage of gaps is far too great. Another observable difference is the steeper decay presented by the frequency of frequencies for the very low frequency region of the random generated texts. For the rest of the frequency domain, we observe that the frequency of frequencies of the natural text is always above the random text (excluding the gaps), what might be explained as a compensatory effect of holding a large amount of gaps.

It is important to note that, strictly when we talk about "the frequency", an assumption is made about the random process which generate the observed data, we assume it does not change with time nor depend on external events. When describing a language, these conditions are generally not considered, and for that reason it would be more accurate to say "the average frequency of occurrence of a word". We know that the frequency of occurrence of words change with time, for example, the frequency of occurrence of *throve* and *thrived* changed in a 200 years gap causing the regularization of the verb *thrive*. The change in the frequency of occurrence might also be occasioned by external factors, for example, the frequency of occurrence the word *influenza* is influenced by the occurrence of epidemic flues (Michel et al., 2011).

The word frequencies in human communications arrange themselves according to the Zipf's law. This model has shown suitable for different languages (Balasubrahmanyana and Naranan, 1996) with no exception found until now. Although different functions have been proposed for modeling the word frequency relation (Tuldava, 1996), the Zipf's law model still is argued as the best and most general model. When we consider the frequency of frequency, we also observed a Zipf's law present in the Natural text. If $P(f)$ is the probability of words with a given frequency f in a corpus, we will observer a relation

$$P(f) \propto f^{-\beta} \tag{7.32}$$

This exponent value is characteristic of the process under analysis. The relation in Equation 7.32 might be interpreted as the frequency spectrum, the value of $P(f)$ refers to the power spectral density related to the frequency f. The value of $\beta = 0$ characterize a white noise process ($1/f^0$); when $\beta = 1$ the process is characterized as a pink noise ($1/f$); and

the value of $\beta = 2$ characterize a Brownian noise or red noise process ($1/f^2$). Usually the value of β approaches 1, characterizing a pink noise process (Mandelbrot, 1999).

The Zipf'law relation presented in Equation 6.2, states that $f \propto k^{-s}$, and from this proportion we may also state that $k \propto f^{-1/s}$. From Equation 7.32, we may find another relation between rank and probability of occurrence of a word. The number of words with a population f in the sample is given by

$$m_f = TP(f) \ , \tag{7.33}$$

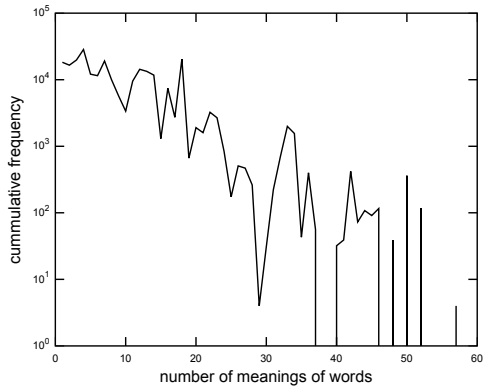
where T is the total number of words in the sample. The rank a words with f observations in the sample is given by

$$\begin{aligned} k(f) &= \int_f^\infty m_{f'} df' \\ &= \int_f^\infty TP(f') df' \\ &= TC_p \int_f^\infty f'^{-\beta} df' \\ &= TC_p f'^{(1-\beta)} \Big|_f^\infty \ , \end{aligned} \tag{7.34}$$

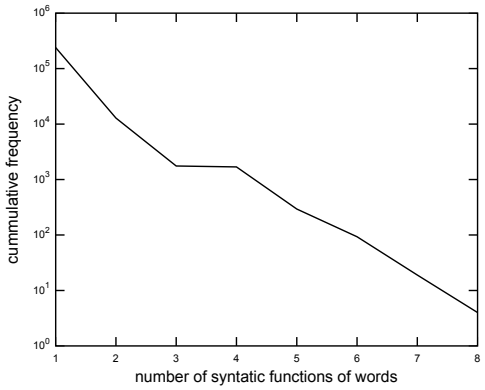
where C_p is the proportionality constant between $P(f)$ and f. From Equation 7.34 we observe that $k \propto f^{(1-\beta)}$ and we may conclude that $f^{-1/s} \approx f^{(1-\beta)}$. There is then a relationship between the exponents

$$s = \frac{1}{\beta - 1} \quad \text{or} \quad \beta = \frac{1}{s} + 1 \ . \tag{7.35}$$

Natural languages typically present a value of $\beta \approx 2$ (or $s \approx 1$ by the relation above), but significant deviations from this typical value have been reported in different contexts. Piotrovskii et al. (1994) after analyzing various samples of schizophrenic language have shown that a $\beta > 2$ is found in the fragmented discourse observed in schizophrenia. The speech is marked by a multitude of topics with no consistent subject, resulting in a varied and chaotic lexicon. In advanced forms of schizophrenia, Piotrovskii et al. (1994) have found an exponent $1 < \beta < 2$. The patients that suffer from this advanced form usually present obsessional topics and the utterance construction is usually filled by words and compounds related to the topic. Very young children present an exponent $\beta \approx 1.6$ (Piotrovskii et al., 1994) and older children conform to the typical $\beta \approx 2$ (Zipf, 1942). Kolguškin (1960) showed that military combat texts present an exponent of $\beta = 1.7$. According to Piotrovskii et al. (1994), a larger value of the exponent $\beta \approx 2$ may be



(a) Semantics: Cumulative frequency of occurrence of words with a given number of meanings.



(b) Syntax: Cumulative frequency of occurrence of words with a given number of syntactic functions.

Figure 7.21: For both pictures the frequency of occurrence data comes from the text *Ulysses*. They present the cumulative frequency of occurrence of words for a certain number of meanings and syntactic functions.

obtained as a result of deficient sampling from a text with the typical $\beta \approx 2$.

Ferrer-i-Cancho (2005b) hypothesized that the variation observed in the exponent value β "reflects our ability to balance the goal of communication, i.e. maximizing the information transfer and the cost of communication, imposed by the limitations of the human brain". The exponent value seems to be related to the communication efficiency, increasing the β value leads to an increase in communicative efficiency. "This positive correlation is not easy to determine, because precise information theory measures, as far as we know, have not been used for the atypical systems considered here" (Ferrer-i-Cancho, 2005b).

7.7 Smoothing

The estimation of the probability of a word given in Equation 7.31 is the MLE. The problem with the MLE is that it predicts that the probability of a word not seen in the corpus is zero. That might be a problem when trying to use the counts in one corpus to estimate what will be seen in another corpus. A language with lots of rare words might suffer with this, especially when the selected corpus does not comprise a fraction of these words. In order to take into account the existence of theses words that were not present in the corpus at hand, we need to make some considerations. A common approach is to add small positive quantities to all events, including the unobserved events. This technique was advocated by Lidstone (1920); Johnson (1932); Jeffreys (1939). When this additive method is applied adding one to every event, is known as the *Add-One* estimator. This

is an obvious and simple approach, which was proposed by Laplace (1902), but it lacks a principled justification and may lead to inaccurate estimates. Gale and Church (1994) investigated the *Add-One* in detail and concluded that it may give approximately accurate estimates only for data-set which obey certain quite implausible numerical constraints: "for Add-One to produce reasonable estimates, it is necessary that the ratio of unseen types to observed types and the ratio of all types to the training sample size be equal. Since there is no reason for a relationship between sample size and the population surveyed, this condition is usually invalid" (Gale and Church, 1994).

A better approach was worked out by Alan Turing and his statistical assistant Irving John Good during their effort in the Second World War to crack the German ciphers for the Enigma machine. Their approach was theoretically well-founded and are proven to perform well. The Good-Turing estimator (Good, 1953) considers that the unseen events together have a probability equal to the sum of the probabilities of all events that were observed only once in the corpus, for they are equally rare, and the fact that one was present in the corpus and the other not is just a mere question of chance. The method proposed by Good (1953) results from an empirical use of Bayes formalism and it can be obtained by significantly different statistical methods. Three examples of derivation of Good's formula are compared by Nádas (1985).

These techniques here analyzed are called *smoothing*. They are used to adjust the maximum likelihood estimates of probabilities to achieve a better estimate when there is insufficient data to approximate them accurately. "The name *smoothing* comes from the fact that these techniques tend to make distributions more uniform, by adjusting low probabilities such as zero probabilities upward, and high probabilities downward. Not only do smoothing methods generally prevent zero probabilities, but they also attempt to improve the accuracy of the model as a whole. Whenever a probability is estimated from few counts, smoothing has the potential to significantly improve estimation" (Chen and Goodman, 1998).

Many linguistic phenomena exist in an infinite multitude of different types. Words and sequences of words are essentially infinite in number, i.e, in any finite amount of text of speech sample, we expect to see words and sequences that were not seen before. It is important to have a good statistical estimate of the occurrence of such types in order to increase the performance of systems that perform tasks such as spelling correction, sense disambiguation and machine translation. For good operation of those systems it is important that the probability of occurrence of unseen events are not estimated as zero.

The frequency of occurrence of words with a given frequency also present a Zipf-like relation. There are few very frequent words, and their frequency of occurrence differ in great magnitude. On the other hand, there are lots of words which happen just once or

twice in a language corpus. If we plot the relation between the number of words that occur with a certain frequency of occurrence, we shall realize that the relation is given by a power law, that might be observed as a straight line in a log-log plot. This type of analysis is also called *inverse Zipf*.

When analyzing the frequency of occurrence of frequencies of words in a language, we have not only to deal with the zero created by those words that never happened, but also the zeros created by the gaps that usually exist in the high frequency region. Suppose that in our toy-corpus the most frequent word *the* has a frequency of occurrence equals to 265,470, and the second most frequent word *of* has a frequency of occurrence equals to 187,168. If we count the number of words that occurred 265,470 times in the corpus, it will be only one, *the*. There will be no word which occurred 265,469 times, nor 265,471 times, we might conclude that the number of words which have a frequency of occurrence from 265,469 to 187,169 will be zero, and the same happens for words with a frequency of occurrence above 265,470. It seems reasonable to imagine that the word *the* could have happened 265,471 times or 265,469 times or any other number around the actual value found in the corpus. It is desired then to consider that the probability of occurrence of words with frequency 265,471, 265,469, etc. is not zero. On the other extreme, we shall observe many words that occur only once, and a smaller number of words that occur twice, and so on. Observing the number of words for a given frequency of occurrence, we expect to experience a monotonically decreasing number of words as the frequency of occurrence increases.

In our toy-corpus example, we might suppose that words like *obliteration*, *freckles* and *apotheosis* happen only twice, and words like *headroom*, *smugness* and *hazelwood* happen just once. Adding up all words with just one occurrence, we get the total of 4,812. The observed number of words whose frequency of occurrence is two is 2,345. We shall use the notation N_f to indicate the number of occurrence of words with a given frequency (the frequency of frequency). In our example above, we have $N_{265,470} = 1$, $N_{265,469} = 0$, $N_{265,468} = 0$, $N_{187,168} = 1$, $N_2 = 2,345$ and $N_1 = 4,812$. The total number of tokens observed in our corpus might be calculated by

$$N = \sum_f f N_f \tag{7.36}$$

Good-Turing methodology estimate that the total probability of all unseen events is equal to the sum of probabilities of all events that occur only once ($N_0 = N_1$), that gives the following probability for the unseen events

$$p_0 = \frac{N_1}{N} \tag{7.37}$$

We denote here by p_f the probability of events that are observed f times. Using the MLE the estimate, the events not observed have $p_0 = 0$. As we have stated before, we want methods which give an estimate for p_0 which exceeds zero, as is the case for the Good-Turing estimate, which will give the value $p_0 = N_0/N > 0$.

The Good-Turing method (Good, 1953) defines N_f^* to be such that $p_f \equiv N_f^*/N$. In the MLE, we have $N_f^* = N_f$. The values of N_f^* for $f \geq 1$ must be chosen so that the sum of all probabilities is equals to one:

$$\sum_f p_f = 1 \tag{7.38}$$

We must then reduce the probability of the seen events in order to take into account the probability of the unseen events, in such a form that equation 7.38 is still true. The Good-Turing method states that

$$f^* = (f+1)\frac{E[N_{f+1}]}{E[N_f]} \tag{7.39}$$

where $E[\cdot]$ represents the expectation of a random variable. f^* is usually called the "adjusted number of observations", that represents how many words you are expected to see with a given frequency of occurrence. The probability of the unseen events will be approximated by $E[N_1]/N$. The value of N_1 is the largest value and the best estimate among all others N_f. For that reason, to use the value of N_1 is a good approximation to the value of $E[N_1]$.

The Good-Turing estimate is a central procedure for other smoothing techniques. To derive the estimate proposed in Equation 7.39, let assume there are s different types $\alpha_1, \alpha_2, \ldots, \alpha_s$ and that their underlying probabilities are respectively $p_1, p_2, \ldots, p_s$. We might estimate the probability of a type given we know how many occurrences of that type exists on our sample. It will be expressed by $E[p_i|c(\alpha_i) = f]$, where $E[\cdot]$ denotes the expected value and $c(\alpha_i)$ denotes the number of times the type α_i has occurred on the given data. The expected value given above might be expanded as

$$E[p_i|c(\alpha_i) = f] = \sum_{j=1}^{s} p(i = j|c(\alpha_i) = f)p_j \tag{7.40}$$

where $p(i = j|c(\alpha_i) = f)$ is the probability that the unknown type α_i, with f occurrences in the sample, actually is the jth type with underlying probability p_j. The probability $p(i = j|c(\alpha_i) = f)$ might be written as the probability that the type α_j appears f times

in the data divided by the sum of the probabilities for all types.

$$\begin{aligned} p(i=j|c(\alpha_i)=f) &= \frac{p(c(\alpha_j)=f)}{\sum_{j=1}^{s} p(c(\alpha_j)=f)} \\ &= \frac{\binom{N}{f} p_j^f (1-p_j)^{N-f}}{\sum_{j=1}^{s} \binom{N}{f} p_j^r (1-p_j)^{N-f}} \\ &= \frac{p_j^f (1-p_j)^{N-f}}{\sum_{j=1}^{s} p_j^f (1-p_j)^{N-f}} \end{aligned} \tag{7.41}$$

where N is the total number of counts (tokens) in the sample, what might written as $N = \sum_{j=1}^{s} c(\alpha_j)$.

Substituting the result from Equation 7.41 in Equation 7.44 we get

$$E[p_i|c(\alpha_i)=f] = \frac{\sum_{j=1}^{s} p_j^f (1-p_j)^{N-f}}{\sum_{j=1}^{s} p_j^f (1-p_j)^{N-f}} \ . \tag{7.42}$$

Consider now $E_N[n_f]$ the expected number of types which present f counts in a sample of size N. This is equal to the sum of the probability that each type has exactly f counts

$$E_N[n_f] = \sum_{j=1}^{s} p(c(\alpha_j)=f) = \sum_{j=1}^{s} \binom{N}{f} p_j^f (1-p_j)^{N-f} \ . \tag{7.43}$$

Using Equation 7.43 in Equation 7.44 we may write

$$E[p_i|c(\alpha_i)=f] = \frac{f+1}{N+1} \frac{E_{N+1}[n_{f+1}]}{E_N[n_f]} \ . \tag{7.44}$$

We have just found an estimate for the expected probability of a type α_i with f counts. The Good-Turing probability based on the correct count value is given by

$$p_{GT}(\alpha) = \frac{f^*}{N} \ . \tag{7.45}$$

Using Equation 7.45 in conjunction with Equation 7.44 we may write the corrected counting value as

$$f^* = N \frac{f+1}{N+1} \frac{E_{N+1}[n_{f+1}]}{E_N[n_f]} \approx (f+1) \frac{n_{f+1}}{n_f} \ , \tag{7.46}$$

where we have used the following approximations $N/(N+1) \approx 1$, $E_N[n_f] \approx n_f$ and $E_{N+1}[n_{f+1}] \approx n_{f+1}$. The empirical values of n_f are used to estimate their expected values.

One problem with the Good-Turing estimation is that it cannot be used when $n_f = 0$, what unfortunately is really common for high values of f. Gale and Sampson (1995)

propose a smoothing method, based on Good-Turing, which overcomes this difficulty.

7.7.1 Simple Good-Turing

A simple way to do a Good-Turing estimation (Gale and Sampson, 1995) is to choose a $E[\cdot]$ so that

$$E[N_{f+1}] = E[N_f]\left(\frac{f}{f+1}\right)\left(1 - \frac{E[N_1]}{N}\right) \tag{7.47}$$

Using the relation 7.47 and 7.39, the probability of occurrence of words with a given frequency f will be given by

$$\begin{aligned} p_f^* &= \frac{f^* N_f^*}{N} \\ &= \frac{(f+1)N_f E[N_{f+1}]}{NE[N_f]} \\ &= \frac{(f+1)N_f E[N_f]\frac{f}{f+1}\left(1 - \frac{E[N_1]}{N}\right)}{NE[N_f]} \\ &= \frac{fN_f}{N}\left(1 - \frac{E[N_1]}{N}\right) \\ &= p_f\left(1 - \frac{E[N_1]}{N}\right) \end{aligned} \tag{7.48}$$

It means that the new estimated probability of a given type, if the sample were perfectly representative of the population, is given by the previous probability, which regards only the observed samples, multiplied by a factor to take into account the unseen types. This relation means to scale down the maximum likelihood estimator fN_f/N by a factor of $(1 - E[N_1]/N)$. If we sum all p_f^* for every seen word in the corpus we get

$$\begin{aligned} \sum_f p_f^* &= \sum_f \frac{fN_f}{N}\left(1 - \frac{E[N_1]}{N}\right) \\ &= \frac{1}{N}\left(1 - \frac{E[N_1]}{N}\right)\sum_f fN_f \\ &= \left(1 - \frac{E[N_1]}{N}\right) \end{aligned} \tag{7.49}$$

since $\sum_f N_f$ for the seen words is exactly N. The result given above agrees with what we have specified before, that the probability of unseen words would be E_1/N, adding the probability of all words (seen and unseen) we get 1 as result.

In the method propose in Equation 7.47, we need the value of $f + 1$ to make the new estimation $E[N_{f+1}]$. Unfortunately, as we move forward increasing the value of f we are

going to find some gaps, and large gaps in the region of high values of f. Those gaps are zeros, and for that reason the equation 7.47 should not be applied to estate those estimates. A modified version of the presented method, called the Simple Good-Turing method (SGT) (Gale, 1994), states that, for these point where zeros were found, we should use instead the best fit power law to approximate these values of f.

In order to do so, a new variable Z_f is defined as

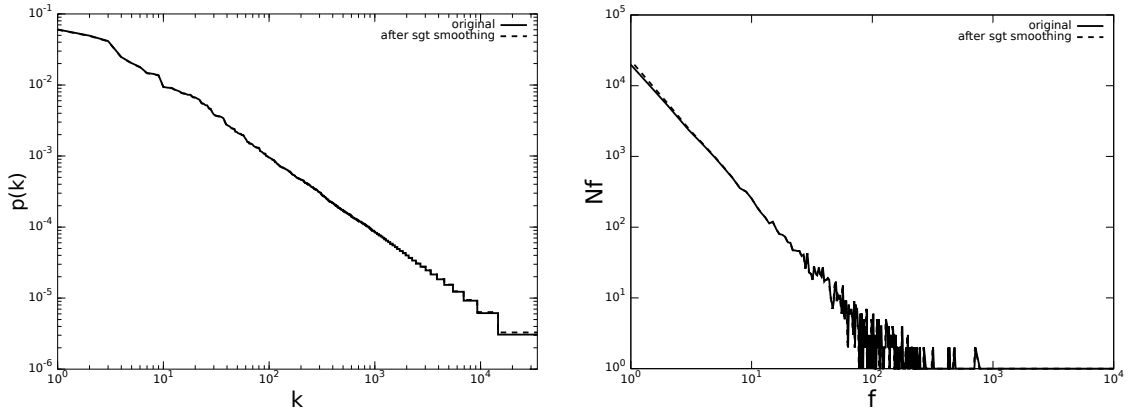
$$Z_f = \frac{2N_f}{f'' - f'} \tag{7.50}$$

where f' is the nearest lower sample frequency and f'' is the nearest higher sample frequency such that $N_{f'}$ and $N_{f''}$ are both nonzero. The log-log plot of f and Z_f typically shows a linear trend, and for that reason a straight line is used as the simplest possible smoothing. A least squared error method is then used to fit the best line. A criteria is used to select whether to use the new approximation f^* or the liner fit approximation. This criteria is based on the standard deviation of the estimate based on N_f. The pair of f^* estimates may be considered significantly different if their difference exceeds 1.96 times the standard deviation (the square root of the variance). "Assuming a Gaussian distribution of the estimate, the probability of such a difference occurring by chance is less than the accepted .05 significance criteria. (...) It is the adoption of a rule for switching between smoothed and raw frequencies of frequencies which allows the SGT method to use such a simple smoothing technique. Good-Turing methods described previously have relied on smoothed proxies for all values of f, and this has forced them to use smoothing calculations which are far more daunting than that of SGT" (Gale and Sampson, 1995). The variance for the Turing estimate is approximately

$$Var(f_T^*) \approx (f+1)^2 \frac{N_{f+1}}{N_f^2} \left(1 + \frac{N_{f+1}}{N_f}\right) \tag{7.51}$$

Any method of smoothing data must satisfy certain prior expectations about f^*. First we expect that f^* will be less than f, for all nonzero values of f; and second, we expect the ratio f^*/f to approach unity as f increases. An example is shown in Figure 7.22 where the corpus used was the text *Ulysses* by James Joyce. In this example f is the frequency of occurrence of words and N_f is the number of words (types) that present f occurrences (tokens) in the corpus. We might observe that the smoothed version present a smaller value of f^* and the ratio f^*/f approaches unity as f increases, since the larger f is, the better it is measured, and for that reason f^* should be closer to f, compared to lower values of f.

The smoothing procedure defined by the SGT method presuppose that all unseen



(a) Zipf plot (b) Inverse Zipf plot

Figure 7.22: Simple Good-Turing Smoothing applied to Ulysses data (words).

objects together amount a frequency of occurrence of those objects seen once. If we decide to divide this probability equally amount the unseen objects, it requires to rely on an assumption over the structure of the objects, and that decision might depend on the application at hand. Church and Gale (1991) show one example using bigram of pair of words where they use two alternative methods, one of them is an enhanced version of the Good-Turing method using "the modest assumption that the distribution of each bigram is binomial"(Church and Gale, 1991). The methods analyzed use a second predictor of the probability in addition to the observed frequency, making it possible to estimate different probabilities for bigrams with the same frequency (that is the reason the method is known as an *enhanced* Good-Turing method).

Samuelsson (1996) presents the relation between Turing's smoothing formula and Zipf's law, using an asymptotic approximation for population frequencies derived from Turing's formula and a local reestimation formula derived from Zipf's law. The two are shown to be instances from a common class of reestimation-formula, although they are qualitatively different. The Turing's formula is shown to "smooths the frequency estimates towards a geometric distribution. (...) Although the two equations are similar, Turing's formula shifts the frequency mass towards more frequent species.".

7.8 Information

The main concern in communication is delivering a message containing information. This process always involves a sender and a receiver, and it requires certain agreements between the parts. If the message is misunderstood at the receiving end, the communication process failures. In every communication the transmitted information is *a prior* unknown, and therefore there might always be doubt if the information extracted from

the received message corresponds to the correct information the sender meant to send. The study of the communication process and the effective transmission of information is the main motivation in Information Theory.

In a communication process there are at least two participants which are physically apart. The information must somehow pass through the surrounding medium to achieve its goal. For that reason it must be modulated appropriately. The modulation used is intrinsically linked to the medium and its properties. The modulated message is an representation of the primary information into another form which is suitable to the communication process across the surround medium. In a speech communication process, the speaker is the source that produces a message which is intended to be received by a listener. The information to be sent is modulated into an utterance which travels through the air medium as a acoustic wave and is further perceived by the listener. To achieve communication successfully, it is necessary that both speaker and listener share the same code.

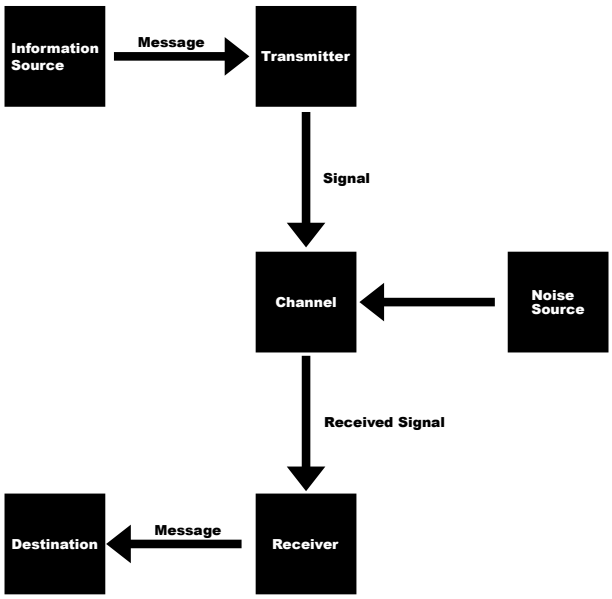


Figure 7.23: Schematics of a Communication System

The information is coded through signs, which are representations of symbols. The study of information by the *Information Theory* has a postulate that it might be represented by a set or sequence of symbols, and symbols are discrete in nature. What is perceived by the receptor are just signs of the coded information. The discreteness feature is very important, it makes the whole system operates in a discrete set and that makes it possible to undertake corrections when a signal is distorted. The communication is always degraded by environmental noise, so distortion is inevitable. To achieve good communication it is desired that system has correction capabilities.

If the set of symbols is finite, and also the duration of any message, we may conclude that each message may be represented by a finite subset of symbols, and coded in a finite time. The cardinality of the set is determined by the coding process, the symbols used and the complexity of the information conveyed.

One coded message to be transmitted through a communication system is one selected from a set of many possible combinations of symbols. Concerning a measure of the information in a message, Shannon pointed out: "If the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely"(Shannon, 1948). The use of a logarithmic function for measuring information was proposed by Hartley (Hartley, 1928) in 1928 and is more convenient for various reasons: linear variation with the logarithm of the number of possibilities; resemblance to our intuitive feeling of a measure of information (two memory cards have twice the capacity of information storage of one memory card; and two communication channels have twice the capacity of transmitting information compared to one single channel); it is mathematically more suitable.

Figure 7.23 depicts a schematics of a generic communication system. This representation comprises 6 basic blocks. The 'information source' is the one which produces a message to be transmitted. This message, which conveys information, may be of many sorts, according to the problem at hand. If we are analyzing a telegraph system, we consider a sequence of letters as the message to be transmitted; in a speech conversation, we may consider a sequence of words, representing the message that conveys information; in a radio transmission, the message is a function of time $f(t)$ which express the acoustic signal that carries music or speech content. The second block is the 'transmitter', which is responsible for coding the input message into some sort of signal that will be suitable for transmission over the channel. In the telegraphic example, the transmitter is responsible in converting the letters sequence into a sequence of dots, dashes and spaces. In the speech communication, the transmitter is responsible for coding and producing a signal by means of articulation of the oral tract. In the radio transmission example, the transmitter performs the FM modulation on the signal, so that the information of $f(t)$ will now be represented by a modulation in frequency of a carrier. The next stage in the communication system is the channel. At this stage we may not control all the pertaining conditions as desired to a successful communication. The channel is indeed the medium where the signal will be transmitted. It may be a pair of wires, as in the telegraphic example, or the free air, as in the speech communication example and the radio broadcast example. The channel is then susceptible to noise, that means the signal may be corrupted by other undesirable sources. The pair of wires is susceptible to induced

electrical field from other sources in the surrounding, it is susceptible to thermal noise, caused by natural random movements of the free electrons in a conductor. The free air is susceptible to climatic variations and also to many other sources that also use this medium for communication, in the speech communication example we have the popular 'cocktail party' example. The transmitted signal travels all the medium from the transmitter until reaching the 'receiver', as a final destination. The receiver has to do exactly the opposite done by the transmitter. The receiver attempts to reconstruct the transmitted message, performing the decoding process and corrections when necessary. In the 'cocktail party' example, we know from experience that corrections do actually occur. In such loud conditions, we may understand everything someone says although the signal received is heavily corrupted. Finally, the message reaches the destination, where the information received will be interpreted, and the communication process is then finished.

7.8.1 The Measurement of Information

A quantitative measure of information was proposed by (Hartley, 1928), contrasting physical aspects with psychological considerations. The term information is very elastic and it is necessary to define a specific meaning of it. Weather dealing with transmission or storage of information, in our discussion we may consider it in terms of a selection from a group of physical symbols, such as words, phonemes, dots and dashes, which under certain agreement convey certain messages to the communicating parties.

During the communication process, the sender, according to the message (s)he is willing to transmit, selects symbols successively from the symbols inventory. The realization of those symbols is transmitted and received at the other side. The receiver, by means of the receiving signal and through successive selections, has brought to attention a set of symbols, eliminating many other possibilities of what could be the ordered set of symbols received. As the communication process proceeds, more symbols are received, and more possible ordered set of symbols are eliminated, making information more precise.

The precision of information depends not only on what was the set of symbols transmitted, but also on what it could be. For example, consider two sentences and their respective information: 'I bought a red car' and 'I bought a red apple'. The information conveyed by the word 'red' in both sentences is complete disparate. On the first sentence 'red' specify a quality among a great set of possibilities. On the second sentence, there are only two possibilities ('red' or 'green'), so the information added by the word 'red' does not convey as much information as in the first example. Of course, this statement is true only when the communication parties share the same knowledge that apples can only be red or green. We may conclude that the degree of information depends also on the previous understanding on available symbols existing between the communicating parties.

In every communication system, the receiving signal can differ many degrees from the transmitted signal. The corruption of signals is a natural phenomenon every communication system is susceptible to. The traveling signal may be weakened and degraded by noise. When this signal arrives at the receiver it may not be recognized at all or miss-recognized. The ability of the receiver to detect and correct errors in the incoming signal is of great importance to the effective characterization of information.

Considering a system where there are 7 symbols available, the selection of two symbols makes possible 7^2 (or 49) different permutations, the selection of three makes 7^3 (or 343) possibilities and the selection of n makes 7^n. If we have a system with s different symbols available to selection, it makes a total of s^n permutation possibilities when considering a process of n choices in this system. Let us consider now an example where we have two systems with different symbol repertories. The first system has s_1 symbols and the second system s_2 on their repertories, respectively. The second system's work is to gather n_1 symbols emitted by the first system and assign to each set of symbols a new symbol from the second system's repertory. We may conclude that the size of the second system's repertory must be at least the total number of possible permutations of n_1 symbols from the first system, and we may chose the lowest available value, because it has no sense to keep unused symbols in our repertory. The number of symbols in the second system's repertory is $s_2 = s_1^{n_1}$. After a communication process proceeds that the second system has transmitted n_2 symbols, leading to $s_2^{n_2}$ choices in that system. If we consider the system that works on the primary symbols, the number of observed symbols will be $n_2 n_1$ symbols. An amount of $n_2 n_1$ symbols is capable of creating $s_1^{n_2 n_1}$ permutations, when working with the first symbol repertory, what must be equivalent to the number of choices of the system using the second repertory: $s_1^{n_1 n_2} = s_2^{n_2}$.

A system with a repertory of s symbols is capable of creating s^n different sequences of symbols when faced with n choices. If we consider the number of different permutations as a measure of information, we would have an exponentially increasing number. Each new symbols added to a sequence would add much more information than the previous ones. That seems contra-intuitive. It would be a more reasonable measure of information, some sort of measure that is proportional to the number of choices and not to the number of possible permutations. The information associated with n selections is given then by

$$H = Kn \ . \tag{7.52}$$

The constant K depends on the number of symbols s, because the information conveyed by a single choice is a function of the number of possibilities in this choice.

Comparing two systems with s_1 and s_2 number of symbols in inventory, and constants K_1 and K_2, respectively associated with them, we may define both constants in such a

way that the amount of information on both system is the same, $H = K_1 n_1 = K_2 n_2$, when the number of selections associated with each one is the same, $s_1^{n_1} = s_2^{n_2}$. Taking this assumption we conclude that

$$\frac{K_1}{\log s_1} = \frac{K_2}{\log s_2} \ . \tag{7.53}$$

This relation will hold for all values of s only if it is related to K by

$$K = K_0 \log s \ , \tag{7.54}$$

where K_0 is the same for every systems and arbitrary. Using (7.54) in (7.52) we get

$$H = n \log s = \log s^n \ . \tag{7.55}$$

This logarithmic measure of information is quite convenient for many aspects. It is simple and mathematically easily tractable. This view of informations is also in consonance with the law of Weber and Fechner[6].

The information associated with a single selection ($n = 1$) is the logarithm of the number of available symbols. If the process involves n selections, the information associated will be n times greater than the information associated with a single section. The numerical value of information will depend on the base chosen for the logarithm. Considering the logarithm in base 2 and the process of one selection from an inventory of only two symbols, the information associated will be $\log_2 2 = 1$, and we say the information associated is 1 bit (bit is used when working with a base 2). The increase of the number of selections from n to $2n$ will lead to an increase of $\frac{2n \log s}{n \log s} = 2$ times the information associated, and this results holds whatever the logarithm based is used. Using the same reasoning, an increase in the number of symbols in the inventory for a factor of two, from s to $2s$, will lead to an increase of $\frac{n \log 2s}{n \log s} = \frac{\log 2 + \log s}{\log s} = \frac{\log 2}{\log s} + 1$ times the information associated. In this last situation, the degree of increase in the information depends on the number of symbols in the inventory. Considering the situation where $s = 2$, the increase of information would be of 2 times; considering $s = 4$, the increase would be of 1.5 times. It will lead to the same 2 times increase in information only when the number of symbols in the repertory is 2, for all other situation, it will lead to a smaller increase in information.

[6]The law postulated by Weber and Fechner says that, in human perception of physical stimuli, the perceptions are proportional to the logarithm of their stimuli. Consider, for example, the perception of sound of different pitches. The difference of one musical half-tone is given by the increase of $2^{1/12}$ times its frequency (approximately 6%). Another example is the perception of loudness of sounds. The perception of loudness follows a relation of the form $L = 10 \log_{10} S$, where L stands for loudness and S for sound pressure level (W/m^2).

7.8.2 Information and Entropy

The concept of information is rather diffuse and broad to be stated by a single definition. For random variables, governed by probability distribution, a quantity called *entropy* attains many properties of what are considered to be a intuitive notion of what information should be. The entropy of one random variable is also called *self-information* of that random variable. A more general definition of entropy is created, the so called *relative entropy*, what is taken as a measure of mutual information between two random variables. The *relative entropy* is in fact a measure of the distance between two probability distributions.

Entropy is a measure of the uncertainty associated with a random variable. The more uncertain a random variable is, the greater entropy it has; and the more certain, the less entropy. Entropy is then a measure based on the probability distribution $p(x)$ (or $p_X(x)$, for a more rigorous notation) of a random variable X. If we are dealing with a discrete random variable X, that means, the values x of X are taken from an alphabet $\mathcal{X}$, we may say then $x \in \mathcal{X}$. The entropy $H(.)$ of the random variable X is defined by

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \ . \tag{7.56}$$

Note that the measure of entropy is a function of the distribution of X, it does not depend on the actual values taken by X.

The entropy of X as defined in (7.56) may be seen as the expectation of another random variable defined as a function of first, $g(X) = \log \frac{1}{p(X)}$. The entropy is then

$$H(X) = E\left[\log \frac{1}{p(X)}\right] = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} \tag{7.57}$$

Considering a finite alphabet $\mathcal{X}_n = \{x_1, x_2, \ldots, x_n\}$ with respective discrete probability distribution $P = \{p_1, p_2, \ldots, p_n\}$, where p_i is probability associated to x_i ($p_i = p(x_i)$). The entropy of the discrete random variable X is

$$H_n(X) = \mathbb{E}_X[I(x)] = -\sum_{x=1}^{n} p(x_i) \log p(x_i) \ . \tag{7.58}$$

We might also write it as $H_n(p_1, p_2, \ldots, p_n)$. $I(x)$ is the self-information, which is the entropy contribution of an individual message.

From the definition of entropy we can easily show some of its properties:

Continuity As a measure, it is required to have continuity, so that small variations in the values of probabilities leads to small changes in the measure of entropy. By

the definition of entropy it is easily seen that it is continuous in relation to the probabilities.

$$H_n(p_1, p_2 + \delta, \ldots, p_n) = H_n(p_1, p_2, \ldots, p_n) + \Delta \ . \tag{7.59}$$

Symmetry It is required that a measure does not change when the probabilities are reordered

$$H_n(p_1, p_2, \ldots, p_n) = H_n(p_2, p_1, \ldots, p_n) \ . \tag{7.60}$$

Extremal Property The maximum entropy happens when the events are equally likely, what is the highest uncertainty situation.

$$H_n(p_1, p_2, \ldots, p_n) \leq H_n\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right) \ . \tag{7.61}$$

Conversely, once we know which specific event among a number of n equally likely events has occurred, we have acquired the largest average amount of information relevant to the occurrence of events of a universe consisting of n complete events.

Considering equiprobable random variables, the entropy should increase with the number of possible outcomes

$$H_n\Big(\underbrace{\frac{1}{n}, \ldots, \frac{1}{n}}_{n}\Big) < H_{n+1}\Big(\underbrace{\frac{1}{n+1}, \ldots, \frac{1}{n+1}}_{n+1}\Big) \ . \tag{7.62}$$

Additivity Regarded as how a process is divided into parts, the amount of entropy should be independent of it. That means, if we consider a system divided into subsystems, the entropy of the system as a whole can be calculated from the entropies of its subsystems.

Suppose we have a uniform random variable X, which outcomes are taken from $\mathcal{X}$ with cardinality n. This random variable is input to a system which divides the excursion range of X into k intervals. At each i-th interval, there will be b_i outcomes of X in this interval, then we have $b_1 + \ldots + b_k = n$. The entropy of the whole system should be equal to the sum of the entropy of the system of intervals, added by the sum of the individual entropies of each interval weighted by the probability of that particular interval.

$$H_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) = H_k\left(\frac{b_1}{n}, \ldots, \frac{b_k}{n}\right) + \sum_{i=1}^{k} \frac{b_i}{n} H_{b_i}\left(\frac{1}{b_i}, \ldots, \frac{1}{b_i}\right) \ . \tag{7.63}$$

Non Negativity The entropy, as a measure, should have a positive or null value, that

means

$$H(X) \geq 0 \ . \tag{7.64}$$

For the Shannon's definition in (7.56), since $0 \leq p(x) \leq 1$, it implies that $\frac{1}{p(x)} \geq 1$, then $-\log p(x) \geq 0$ and it follows that $H(X) \geq 0$.

Event of Null Probability The events of null probability do not contribute to the entropy.

$$H_{n+1}\left(p_1, p_2, \ldots, p_n, 0\right) = H_n\left(p_1, p_2, \ldots, p_n\right) \ . \tag{7.65}$$

Jensen Inequality The entropy of a random variable X is bounded to the logarithm of the cardinality n of the inventory set of the random variable.

$$H(X) = E\left[\log\left(\frac{1}{p(X)}\right)\right] \leq \log\left[E\left(\frac{1}{p(X)}\right)\right] = \log(n) \ . \tag{7.66}$$

7.8.3 Mutual Information

Mutual information is defined as a measure of the amount of information that may be obtained from one random variable when another is observed. In communication it is important to maximize the amount of information shared between sent and received signals. The mutual information between two random variables X and Y is defined by

$$I(X;Y) = \mathbb{E}_{X,Y}[SI(x,y)] = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)\, p(y)} \tag{7.67}$$

A basic property of the mutual information is that

$$I(X;Y) = H(X) - H(X|Y), \tag{7.68}$$

meaning that the mutual information of X and Y is equals to the difference between the self-information of X and the amount of uncertainty on X given that Y is known.

7.8.4 Data-Processing Inequality

An important theorem in information theory is the theorem of data processing. This theorem states that it is impossible to perform any data processing that will leads to improvement on the inferences possible to be made over this data.

Theorem 7.8.1 (Data-Processing Inequality) *If the random variables X, Y and Z make a Markov chain in this order, $X \to Y \to Z$, then $I(X;Y) \geq I(X;Z)$.*

Proof If the random variables create a Markov chain in this order, that means the conditional distribution of Z is dependent only on Y and conditionally independent of X. The joint probability density function may be written as follows

$$p(x, y, z) = p(x)p(y|x)p(z|y). \tag{7.69}$$

A Markov chain holds and is possible only if X and Z are independents given Y.

$$\begin{aligned} p(x, z|y) &= p(x, y, z)/p(y) \\ &= p(x, y)p(z|y)/p(y) \\ &= p(x|y)p(z|y). \end{aligned} \tag{7.70}$$

Markov chain implies conditional independence.

Considering now the mutual information

$$\begin{aligned} I(X; Y, Z) &= I(X; Z) + I(X; Y|Z) \\ &= I(X; Y) + I(X; Z|Y), \end{aligned} \tag{7.71}$$

and the fact that $I(X; Z|Y) = 0$, since a Markov chain implies that X and Z are conditionally independent given Y; and as $I(X; Y|Z) \geq 0$, we have

$$I(X; Y) \geq I(X; Z). \tag{7.72}$$

The equality holds only if $I(X; Y|Z) = 0$, what means that $X \to Z \to Y$ is also a Markov chain. In a similar way, we conclude that $I(Y; Z) \geq I(X; Z)$.

If we consider $Z = g(Y)$, we have a Markov chain as $X \to Y \to g(Y)$, then $I(X; Y) \geq I(X; g(Y))$. A function over the data Y is not able to increase the information over X. □

Remark If X is transmitted signal and Y the received signal. If we want to make inferences over X using the known outcomes of Y, it is of no use to make a processing over the data Y creating $Z = g(Y)$, because the resultant Z has a mutual information with X that is equal or less than the mutual information between Y and X.

7.8.5 Conditioning and Entropy

If we consider a set of symbols, it is intuitive to think that the knowledge of many symbols reduces the entropy over one unknown symbol. The context may be used to bring incite on what may be the missing symbol. Considering the coding of a sequence of

symbols, the entropy of the next symbol may be reduced by the knowledge of the previous ones.

Theorem 7.8.2 (conditioning reduces entropy) *For two random variables X and Y, $H(X|Y) \leq H(X)$ and the equality holds only if X and Y are independents.*

Proof The proof of this theorem is using the Jensen's inequality.

$$\begin{aligned} H(X|Y) - H(X) &= \sum_{x\in X, y\in Y} p(x,y) \log \frac{1}{p(x,y)} - \sum_{x\in X} p(x) \log \frac{1}{p(x)} \\ &= \sum_{x\in X, y\in Y} p(x,y) \left(\log \frac{p(y)}{p(x,y)} + \log p(x) \right) \\ &= \sum_{x\in X, y\in Y} p(x,y) \left(\log \frac{p(y)p(x)}{p(x,y)} \right) \\ &\leq \log \sum_{x\in X, y\in Y} p(x,y) \frac{p(y)p(x)}{p(x,y)} \quad \text{(by Jensen's inequality)} \\ &= 0. \end{aligned} \tag{7.73}$$

And we have then $H(X|Y) \leq H(X)$. □

Remark It is important to note that what this theorem means is that *on average* if we know the value of Y then our uncertainty about X is reduced. For a specific value $Y = y$, it is not possible to tell whether $H(X|Y = y)$ is greater, equal or smaller than $H(X)$. That means that the knowledge of one new evidence $Y = y$ may increase the uncertainty over a variable, but on average it will lead to a reduction on the uncertainty.

$$H(X|Y) = \sum_y p(y) H(X|Y = y) \leq H(X). \tag{7.74}$$

7.8.6 Language and Entropy

Several features, present in every language, suggest the existence of a fundamental principle of language organization. The Zipf law, as evidenced previously, is the best known and attested in several languages. Language, as a communication system, is bounded by the physical and biological apparatus and also by the communication system requirements. Our spoken communication process is bounded by the speaker trying to encode a message, and by a listener trying to decode the received signal. A trade-off exists between the sender and the receiver. The speaker wants to minimize articulatory effort, creating a tendency for brevity and phonological reduction. On the other side, the hearer want to minimize the understanding effort, desiring explicitness and clarity. The hearer

wants to avoid ambiguities and prefer the most dissimilar sound clusters. The availability of words is positively correlated with their frequency of occurrence and, for the speaker, it is preferable to use the most frequent words, tending to choose the most ambiguous words.

Mandelbrot (1953) dealt with the study of Language Statistical Structure under the point of view of Information Theory. He pointed out that "this problem is an inverse of the classical problem of C. E. Shannon, that is to say, of the *direct* problem of constructing the least costly coding for a given message". When analyzing language under the informational theory point of view, we shall assume that every communication assumes the existence of two participants, a speaker and a listener.

The Shannon information entropy for a given source with a set with R distinct symbols will have its entropy defined by

$$H_R = -\sum_{i=0}^{R-1} p_i log_R p_i \tag{7.75}$$

where p_i stands for the probability of the i-th symbol in the set and R is the set length, which is chosen for the base of the logarithm function because in that manner we may normalize the entropy in the range $(0, 1)$ what is useful since we want to compare sources with different symbol sets.

Natural languages are expressed as a symbolic sequence that shows a balance between order and disorder, what is determined by the interplay of diversity of symbols and structural constraints in the way they may be organized. Both aspects are important to achieve a language purpose: to convey meaning and communicate. Since the work of Shannon (1948) the problem of assigning an entropy for languages has been of interest for several researches. It is important to bear in mind that linguistics structures are present in various levels on the organization and elaboration of languages. As we consider an entropy measure, whether taking words as our symbolic unities, whether using phones, diphones, triphones or syllables, the contributions of each level of organization to the final entropy measured is already embedded.

Using the definition on equation 7.75, we might calculate the word entropy on natural and random generated texts. Those data present different repertories, with different length (number of elements), and for that reason we need to normalize the entropy measure in order to make it possible to establish comparisons. As it is proposed in equation 7.75, the base R of the logarithm function is the number of elements in the set. The entropy measure obtained for the natural text, the random text using Markov model, the random text with weighted probability of occurrence of the symbols and the white random text are respectively as follows: 0.71089, 0.78346, 0.85418, 0.99509. Unities are omitted due

to the normalization into the interval $(0, 1)$. As the entropy approaches 1, it approaches the maximal entropy, as the source model approaches the ideal white random model. As we might observe by the calculated entropy numbers is that deeper consideration on the statistical models of language, leads to an entropy closer to the natural language entropy.

Zipf frequency-rank and entropy relation doesn't regard any information about the way the symbols are ordered in an utterance or in a written text. You shall scramble the observed data and still the Zipf and entropy relations will hold exactly the same. There is then a certain organization in language that is regarded by the order in which the symbols succeed one another. Language has a long-range correlation making it not possible to accurately compute the entropy of the source by estimating block probabilities directly. The compression algorithm proposed by Ziv and Lempel (1978) is known to compress any stationary and ergodic source to the entropy rate of the source per symbol, provided the input sequence is sufficiently long Cover and Thomas (1991). Montemurro and Zanette (2011) used the efficient entropy estimator derived from the Lempel-Ziv compression algorithm and compares the entropy estimation of a natural text with a randomly sorted version of the original text. They compare the results in different language, from different language families and conclude that "beyond the apparent diversity found between languages, the impact of word ordering stands as a robust universal statistical feature across linguistic families"(Montemurro and Zanette, 2011).

7.8.7 Data Information Content

The information content H in a certain data might be defined in terms of the extent to which each new symbol in a data stream removes uncertainty on the receiver of the given message. It is therefore dependent on the receiver's prior knowledge and on the received message itself. H must be a decreasing function of the a priori probability p with which the receiver could predict the message. If the receiver is able to predict the received message with certainty, then we have $p = 1$ and the message carries no information at all. If the probability p is reduced, the receiver's a priori uncertainty is increased. In the extreme situation, when $p = 0$, the receiver's a priori ignorance may be regarded as infinitely large, and therefore the received message provides the receiver with an infinite amount of information content.

If we consider that the sequence of symbols in a message is a sequence of independent events, it is desirable to have an information content measure that assign an information value to the sequence of symbols equals to the sum of information provided by each symbol independently. That means we seek an information measure H such that

$$H(p_1 p_2) = H(p_1) + H(p_2) \ . \tag{7.76}$$

We are looking for a decreasing function of p, that satisfies Equation 7.76 and $H(1) = 0$. H should also be continuous on p, for small variations on the value of p should lead to small variations on the value of $H(p)$. It is shown (Shannon, 1948) that the only solution to these restrictions is the logarithmic function

$$H(p) = -\log p \ . \tag{7.77}$$

Each symbol in a message with probability p will be said to reduce the receiver's uncertainty, or entropy, by $-\log p$. The base of the logarithm usually used is 2 and therefore the unity of information content used is bits.

Suppose in a communication process we have a vocabulary of N symbols with given probabilities p_i for each i-th symbol in the set. The expected information content to be gained as a new symbol arrives at the receiver is given by

$$\overline{H} = -\sum_{i=1}^{N} p_i \log p_i \ . \tag{7.78}$$

Given the expected information content above, we might wonder what are the probabilities of symbols occurrence that leads to the maximum expected information content. The probabilities must satisfy

$$\sum_{i=1}^{N} p_i = 1 \tag{7.79}$$

and we might rewrite Equation 7.78 taking one term out of the sum, for example, the last one, only for notational simplicity,

$$\overline{H} = -\sum_{i=1}^{N-1} p_i \log p_i - p_N \log p_N \tag{7.80}$$

In the same way, we may express p_N as follows, using Equation 7.79,

$$p_N = 1 - \sum_{i=1}^{N-1} p_i \tag{7.81}$$

$\overline{H}$ is regarded as a function of the $N-1$ probabilities p_i and its maximum value will be when the following condition is satisfied

$$\frac{\partial \overline{H}}{\partial p_k} = 0 \quad \text{for} \quad k = 1, \ldots, N-1 \ . \tag{7.82}$$

In order to simplify our mathematical notation, lets rewrite Equation 7.80 in the

natural logarithm base

$$\overline{H} = -\left(\frac{1}{\ln 2}\right)\left[\sum_{i=1}^{N-1} p_i \ln p_i + p_N \ln p_N\right] \tag{7.83}$$

We might write Equation 7.82 as

$$\begin{aligned} 0 = \frac{\partial \overline{H}}{\partial p_k} &= -\left(\frac{1}{\ln 2}\right)\frac{\partial}{\partial p_k}\left[\sum_{i=1}^{N-1} p_i \ln p_i + p_N \ln p_N\right] \\ &= -\left(\frac{1}{\ln 2}\right)\left[\ln p_k + 1 + (\ln p_N + 1)\frac{\partial p_N}{\partial p_k}\right] \\ &= -\left(\frac{1}{\ln 2}\right)[\ln p_k + 1 - (\ln p_N + 1)] \end{aligned} \tag{7.84}$$

since, from Equation 7.81, we have $\partial p_N/\partial p_k = -1$. Therefore, Equation 7.84 tells we want to find

$$\ln p_k = \ln p_N \tag{7.85}$$

for each k. All those $N-1$ equations will be satisfied when all p_k are equal to $1/N$.

We shall compute the second derivative to make sure we have found a maximum point. From Equation 7.84, we have

$$\begin{aligned} \frac{\partial^2 \overline{H}}{\partial p_k^2} &= -\left(\frac{1}{\ln 2}\right)\frac{\partial}{\partial p_k}[\ln p_k - \ln p_N] \\ &= -\left(\frac{1}{\ln 2}\right)\left[\frac{1}{p_k} + \frac{1}{p_N}\right] \leq 0 \ , \end{aligned} \tag{7.86}$$

since the propabilities are positive numbers. Therefore, selecting $p_k = 1/N$ we find the maximum entropy.

The maximum entropy will be found when each symbol has the same *a priori* probability of occurrence. When a receiver makes no assumption on the characteristics of the message generating source, it must consider equal probabilities among its symbols, since any other assumption implies less uncertainty.

Shannon (1951) has investigated the information content of written English. According to his studies, the average entropy of each letter is $H_1 = -\sum_{i=1}^{26} p_i \log p_i = 4.14$ bits, which is smaller than the maximum entropy given by the uniform distributed probabilities, that would give us an average entropy of $H_0 = -\log(1/26) = 4.70$ bits. If the receiver knows a priori the distribution of letters in the written language, we might conclude that the information content of each letter is reduced by $4.70 - 4.14 = 0.56$ bits.

In a continuous text, we know that the probability of occurrence of a given letter might

be different regarding its previous letter. If the probabilities of occurrence of a character pairs are known p_{ij}, we might estimate the uncertainty involved in the prediction of the j-th character, given the previous i-th character

$$-\log(p_{ij}/p_i) = -\log p_{ij} + \log p_i \tag{7.87}$$

Using this result we might estimate the average entropy associated with a single prediction

$$\overline{H_p} = -\sum_{i,j} p_{ij} \log p_{ij} + \sum_i p_i \log p_i \ . \tag{7.88}$$

When computed, it will lead to a smaller value compared to the average entropy based only on the probability of occurrence of letters, since we have added the information of bigram statistics, reducing uncertainty in each prediction. Shannon (1951) concluded that by using bigram information the average entropy will be reduced to 3.56 bits, the inclusion of the information on trigrams will reduce even further the uncertainty to 3.3 bits, and it might be reduced bellow 1 bit per character if we consider the statistics of longer strings of text.

Shannon (1951) used the known fact that words are distributed according to a Zipfian distribution and estimated the average entropy in words of written English to be 11.82 bits per word, leading to an average $11.82/4.5 = 2.62$ bits per letter. Grignetti (1964) points some inconsistencies in the results derived by Shannon (1951) and recalculated the entropy and found it to be 9.83 bits per word in printed English.

Following the procedure proposed by Grignetti (1964), we might generalize the result to any Zipfian distribution and find what are the limits on the entropy on a language. The entropy of a system using N symbols is written as

$$\overline{H} = -\sum_{k=1}^{N} p_k \log p_k \ , \tag{7.89}$$

using the natural logarithm it might be rewritten as

$$\overline{H} = -\frac{1}{\ln 2} \sum_{k=1}^{N} p_k \ln p_k \ . \tag{7.90}$$

Considering the symbols to be Zipfianly distributed, we shall consider $p_k = Ck^{-s}$, where the constant $1/C$ is the generalized harmonic number, $C^{-1} = \sum_{n=1}^{N} n^{-s}$. Using it, the

entropy might be written as

$$\begin{aligned}
\overline{H} &= -\frac{1}{\ln 2}\sum_{k=1}^{N} Ck^{-s}\ln(Ck^{-s}) \\
&= -\frac{C}{\ln 2}\sum_{k=1}^{N} k^{-s}(\ln C - s\ln k) \\
&= -\frac{C}{\ln 2}\ln C\sum_{k=1}^{N} k^{-s} + \frac{sC}{\ln 2}\sum_{k=1}^{N} k^{-s}\ln k \\
&= \frac{sC}{\ln 2}\sum_{k=1}^{N}\frac{\ln k}{k^s} - \frac{\ln C}{\ln 2}
\end{aligned} \tag{7.91}$$

The summation expressed in Equation 7.91 might be calculated following the same steps given by Grignetti (1964). The function $f(x) = \frac{\ln x}{x^s}$ is plotted in Figure 7.24 for some values of s greater than one, what is usually found in natural languages. Taking the first derivative of f,

$$f'(x) = \frac{x^{s-1}(1 - s\ln x)}{x^{2s}} = \frac{1 - s\ln x}{x^{s+1}} \tag{7.92}$$

we might conclude that f is a decreasing function of x for $x > e^{1/s}$, what might be verified in the Figure 7.24.

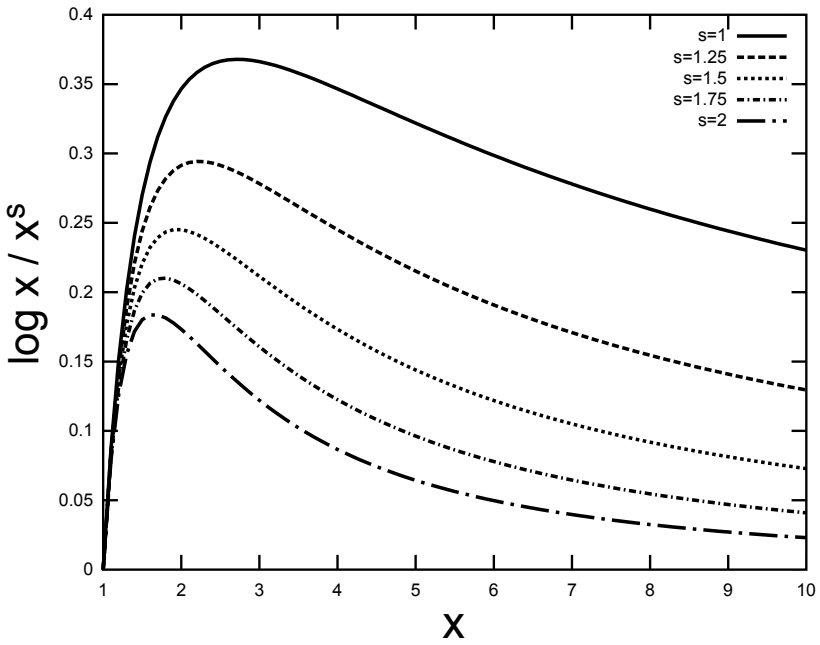


Figure 7.24: Behaviour of the function $f(x) = \frac{\ln x}{x^s}$ for different values of s.

The summation in Equation 7.91, for $x > 3$ is over a decreasing function, regardless which s we are considering. For that reason, we might approximate using the Riemann

sum approximation of an integral. This idea is illustrated in Figure 7.25. As we have a decreasing function, the left Riemann sum is an overestimate and the right Riemann sum is a underestimate.

$$\text{right Riemann sum} \leq \int_a^b f(x)dx \leq \text{left Riemann sum} \tag{7.93}$$

From Equation 7.93 we might write

$$\sum_{n=4}^{N-1} \frac{\ln n}{n^s} \leq \int_3^{N-1} \frac{\ln x}{x^s} dx \leq \sum_{n=3}^{N-2} \frac{\ln n}{n^s} \tag{7.94}$$

and

$$\sum_{n=4}^{N} \frac{\ln n}{n^s} \leq \int_3^{N} \frac{\ln x}{x^s} dx \leq \sum_{n=3}^{N-1} \frac{\ln n}{n^s} \tag{7.95}$$

Using both Equations 7.94 and 7.95 we conclude that

$$\int_3^{N} \frac{\ln x}{x^s} dx \leq \sum_{n=3}^{N-1} \frac{\ln n}{n^s} \leq \int_3^{N-1} \frac{\ln x}{x^s} dx + \frac{\ln 3}{3^s} \tag{7.96}$$

and by adding the two remaining terms to the summation, we get

$$\int_3^{N} \frac{\ln x}{x^s} dx + \frac{\ln 2}{2^s} + \frac{\ln N}{N^s} \leq \sum_{n=1}^{N} \frac{\ln n}{n^s} \leq \int_3^{N-1} \frac{\ln x}{x^s} dx + \frac{\ln 3}{3^s} + \frac{\ln 2}{2^s} + \frac{\ln N}{N^s} \tag{7.97}$$

The integral in Equation 7.97 might be solved using an integration by parts procedure.

$$\int \frac{\ln x}{x^s} dx = \int u dv = uv - \int v du \tag{7.98}$$

We shall choose

$$\begin{aligned} u &= \ln x & dv &= \frac{1}{x^s} dx \\ du &= \frac{1}{x} dx & v &= \frac{x^{-s+1}}{1-s} \end{aligned} \tag{7.99}$$

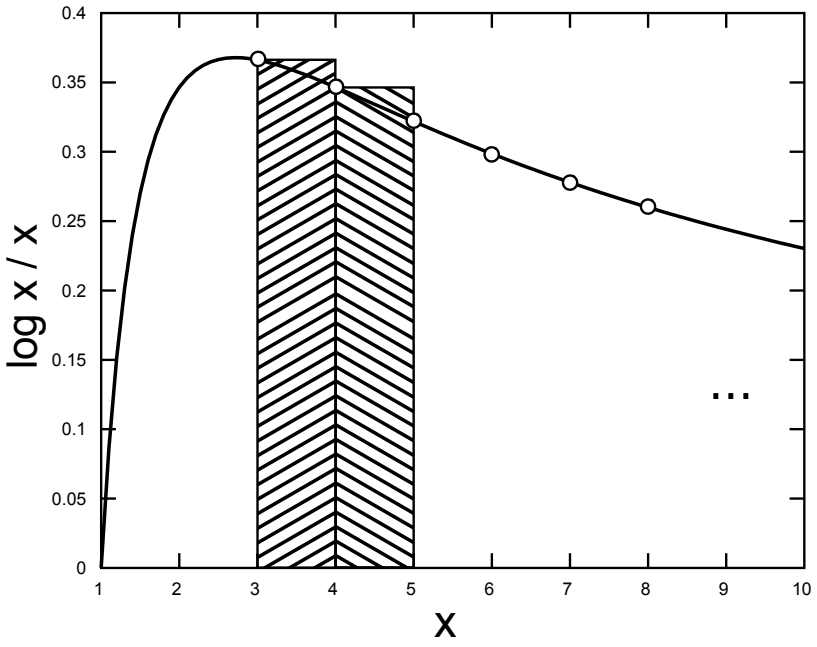


Figure 7.25: Riemann sum approximation of the integral.

Solving the integral, for $s \neq 1$, we have

$$\begin{aligned}
\int \frac{\ln x}{x^s} dx &= \frac{x^{1-s}}{1-s} \ln x - \int \frac{x^{1-s}}{1-s} \frac{1}{x} dx \\
&= \frac{x^{1-s}}{1-s} \ln x - \frac{1}{1-s} \int x^{-s} dx \\
&= \frac{x^{1-s}}{1-s} \ln x - \frac{x^{1-s}}{(1-s)^2} + c \\
&= \frac{x^{1-s}}{1-s} \left(\ln x - \frac{1}{1-s} \right) - c. \qquad (7.100)
\end{aligned}$$

When $s = 1$ the integral will result in

$$\int \frac{\ln x}{x} dx = \frac{(\ln x)^2}{2} + c \; . \qquad (7.101)$$

Using Equations 7.91, 7.97 and 7.100 we are able to calculate the bounds of the entropy of a Zipfian distributed source for given s and N. Figure 7.26 presents some results, where the value of the entropy is given by the average of the upper and lower bounds given by Equation 7.97 in conjunction with Equation 7.91. The lower graph in Figure 7.26 presents the width of the interval to which the estimated entropy is bounded. We might observe that the entropy decreases with s, increases with N and saturates as s decreases towards zero.

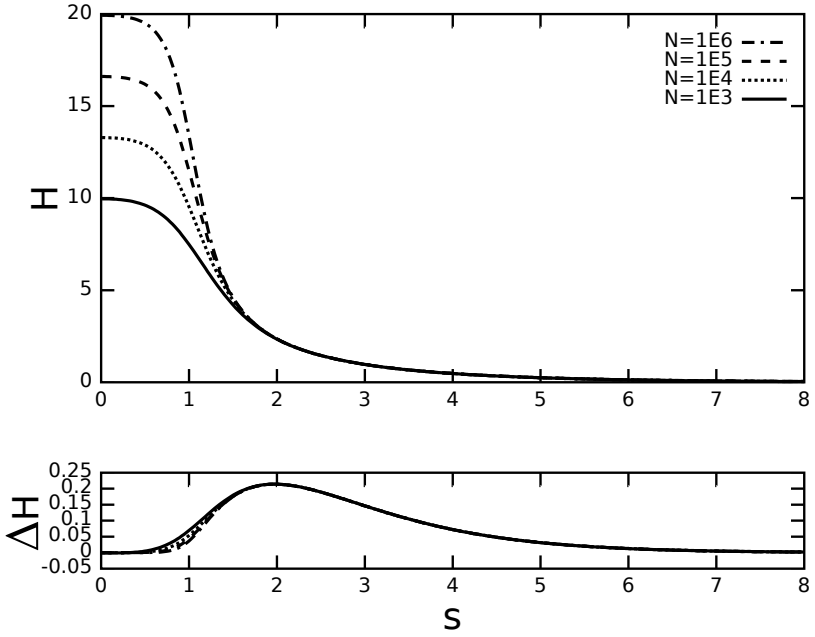


Figure 7.26: Entropy H (given in bits) as a function of the Zipf exponent s and the number of types N. The upper plot presents the average Entropy given by the estimated presented and the lower plot gives the difference between the lower and upper bounds in the estimation.

Entropy of Real Texts

In this section we compare the results from the estimation proposed above with values of entropy computed from real texts. In order to state such comparisons we are going to use a few text corpora: the top 100 most downloaded books in the Gutenberg database; the complete work of William Shakespeare; and Ulysses by James Joice. We used GNU utilities to parse and transform these texts and then count the frequency of occurrence of words in each corpus, creating a word-frequency data-table. This data is afterwards used to compute the entropy of texts, which are going to be compared to the estimated values in Table 7.1 bellow.

Table 7.1: Entropy of real texts (bits) compared with the estimated entropy (bits) using the parameter N (number of types) found in the text and parameter s (Zipf exponent) found by Maximum Likelihood Estimation (MLE).

source	real entropy	N	s_{MLE}	estimated entropy
gutenberg	10.45	142515	1.14	9.17
shakespeare	10.02	27172	1.12	8.68
ulysses	10.63	29994	1.07	9.46

We might observe that there is a significant deviation from the value estimated from the theoretical procedures and the real value obtained from the corpora. Notice that the

estimated value is consistently smaller. This deviation might be explained by a conjunction of discrepancies from our model and the distribution of types found in each corpus. First, we know that in a real rank-frequency plot there is a flattening in the low rank region. This flattening partially responsible for increasing the entropy in the real data. This will we analyzed in the next section when we consider the generalized Zipf-Mandelbrot law (Mandelbrot, 1965). Another aspect responsible for a higher entropy in our corpus is the staircase pattern in high rank region. The existence of many types with the same frequency of occurrence proportionates a higher entropy in comparison with a scenario where these types have distinct probability values. One third aspect that leads to a higher entropy on our corpus is the fact that the probabilities estimated were based on the maximum likelihood. As we pointed out before, it is necessary to perform a smoothing in order to take into account these types which were not present in our sample. A Simple-Good Turing (Gale and Sampson, 1995) could be performed, what would shift the probability mass, decreasing the entropy value.

7.8.8 Zipf-Mandelbrot's Entropy

In order to take into account the flattening observed on low rank region of a Zipf plot, Mandelbrot (1965) introduced a modification on the Zipf's law, adding a constant q to the rank k, resulting in the Zipf-Mandelbrot's law: $p_k(s, q, N) = C(k + q)^{-s}$, where the new normalizing constant (a generalization of a harmonic number) is given by $C^{-1} = \sum_{n=1}^{N}(n + q)^{-s}$.

Applying the same steps to this generalized formulation, the entropy will be given by

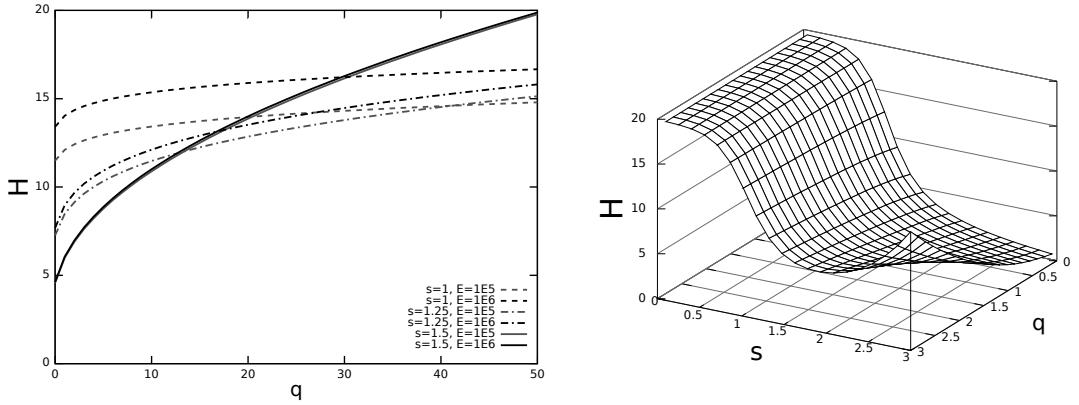
$$\overline{H} = \frac{sC}{\ln 2}\sum_{k=1}^{N}\frac{\ln(k+q)}{(k+q)^s} - \frac{\ln C}{\ln 2}\ . \tag{7.102}$$

The new function $f(x) = (x + q)^{-s}\ln(x + q)$ will be decreasing for $x > e^{1/s} - q$. We shall then define an integer constant $K = \max(\lceil e^{1/s} - q\rceil, 1)$ which guarantees that the function $f(x)$ for $x > K \geq 1$.

Using the left and right Riemann sum again we find the inequalities bellow, which are respectively equivalent to equations 7.94 and 7.95:

$$\sum_{n=K+1}^{N-1}\frac{\ln(n+q)}{(n+q)^s} \leq \int_{K}^{N-1}\frac{\ln(x+q)}{(x+q)^s}dx \leq \sum_{n=K}^{N-2}\frac{\ln(n+q)}{(n+q)^s} \tag{7.103}$$

$$\sum_{n=K+1}^{N}\frac{\ln(n+q)}{(n+q)^s} \leq \int_{K}^{N}\frac{\ln(x+q)}{(x+q)^s}dx \leq \sum_{n=K}^{N-1}\frac{\ln(n+q)}{(n+q)^s}\ . \tag{7.104}$$



(a) Effect of the parameter q on the entropy H (in bits) for a Zipf-Mandelbrot distribution.

(b) Compared effect of the parameter s and q when N is fixed at $1E6$.

Figure 7.27: The effect of the parameter q on the entropy. A greater value of q increases the entropy and the increase step is larger when the value of s is bigger.

From the above equations we conclude that

$$\int_K^N \frac{\ln(x+q)}{(x+q)^s}dx \leq \sum_{n=K}^{N-1} \frac{\ln(n+q)}{(n+q)^s} \leq \int_K^{N-1} \frac{\ln(x+q)}{(x+q)^s}dx + \frac{\ln(K+q)}{(K+q)^s} \tag{7.105}$$

what is the equivalent to Equation 7.96. By adding the remaining terms we get the following boundaries

$$\begin{aligned} B_l &= \int_K^N \frac{\ln(x+q)}{(x+q)^s}dx + \sum_{n=1}^{K-1} \frac{\ln(n+q)}{(n+q)^s} + \frac{\ln(N+q)}{(N+q)^s} \\ &\leq \sum_{n=1}^{N} \frac{\ln(n+q)}{(n+q)^s} \\ &\leq \int_K^{N-1} \frac{\ln(x+q)}{(x+q)^s}dx + \sum_{n=1}^{K} \frac{\ln(n+q)}{(n+q)^s} + \frac{\ln(N+q)}{(N+q)^s} = B_u \ . \end{aligned} \tag{7.106}$$

The integral in Equation 7.106 is solved by parts, giving the same results as presented by equations 7.98 and 7.101, considering that we have $x+q$ instead of x. By adding the parameter q, the distribution suffers a flattening on the low rank values and consequently the entropy of the source increases, what might be observed in Figure 7.27a, where it is shown the effect of an increasing q on distributions where s and N are kept constants.

Table 7.3 presents a new comparison between the entropy estimates and the entropy found in real text data. We might observe that the usage of the Zipf-Mandelbrot model improved the estimation of the entropy of real data. In order to achieve a better approximation, we believe it is necessary to take into account the other aspects mentioned

previously.

Table 7.2: Entropy of real texts (bits) compared with the estimated entropy (bits) using the parameter N (number of types) found in the text, parameter s (Zipf exponent) found by a Maximum Likelihood Estimation (MLE) and the parameter q found empirically.

source	real entropy	N	s_{MLE}	q	estimated entropy
gutenberg	10.45	142515	1.14	3	11.07
shakespeare	10.02	27172	1.12	3	10.92
ulysses	10.63	29994	1.07	2	10.64

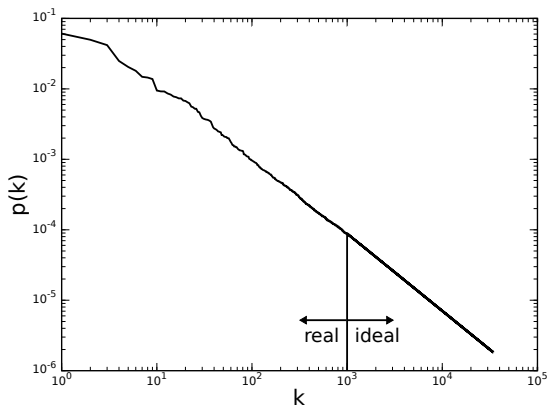
Table 7.3: This table presents the entropy of real texts (bits) with a Simple Good-Turing smoothing applied. They are compared to the estimated entropy (bits) using the parameter N (number of types) found in the text, parameter s (Zipf exponent) found by a Maximum Likelihood Estimation (MLE) and the parameter q found by visual inspection.

source	entropy	after sgt	N	s_{MLE}	q	estimated
carroll alice	8.49	8.79	3016	1.12	2.5	8.78
shakespeare hamlet	9.04	9.08	5447	1.09	1.5	9.07
shakespeare macbeth	9.00	8.76	4017	1.07	1	8.76
shakespeare complete	9.52	9.56	29847	1.15	1.5	9.56
joyce ulysses	10.19	10.25	34391	1.10	1.5	10.24

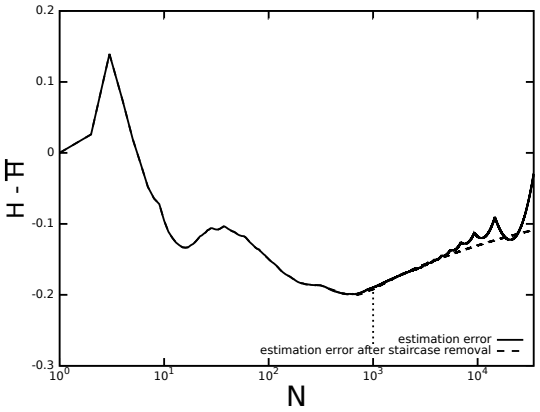
In real texts, rare words appear in a staircase pattern in the end of the long tail of the distribution. It is said to be caused by the undersample these words suffer. In order to investigate the effect of this staircase pattern on the entropy measure we make a substitution on the types with rank over $10E3$ by the corresponding ideal Zipfian model. Figure 7.28a the new probability-rank plot after the substitution. In Figure 7.28b we compare the value of the entropy for the data up to a specific range of rank and the entropy estimated by the method here proposed. We are taking into account the coherence in Zipf (see Section 7.9) and of that reason, a subset of the types up to a certain rank also constitute a set in which the same Zipf's law holds.

7.8.9 Emergence of Zipf's Law

Zipf's law for word frequencies could be the manifestation of a complex system operating between order and disorder. Ferrer-i-Cancho and Solé (2003); Ferrer-i-Cancho (2005a) proposes a model to attest the emergence of the Zipf's law in the balance in communication between information transfer and the cost necessary to achieve communication. It take into compromise the principle of the least effort applied on the listener and speaker, creating a cost function.



(a) Frequency-rank plot where the types with rank over $10E3$ were substituted by an ideal Zipf model to remove the staircase pattern.



(b) Difference between the real entropy and the estimated entropy as a function of N (the number of types).

Figure 7.28: The frequencies of Ulysses types are used to investigate the effect of the staircase pattern on difference between the real entropy and the estimated entropy proposed here.

Every language uses symbolic references and signals to carry information. We might, for example, consider words as our signals, and each one of them make reference to one or more meaning or object in the real world. The model proposed by Ferrer-i-Cancho and Solé (2003); Ferrer-i-Cancho (2005a) assumes there are two finite sets: the set of signals (words) $S = \{s_1, s_2, \ldots, s_n\}$ and the set of stimuli (meanings) $R = \{r_1, r_2, \ldots, r_n\}$. Signals are connected to stimuli and that connections are defined by a binary $n \times m$ matrix $A = \{a_{ij}\}$ where $a_{ij} = 1$ if s_i and r_j are linked and $a_{ij} = 0$ otherwise. The presented links in matrix A don't claim that certain words in S refer to stimuli in R. When an element a_{ij} is equals to one, there is no imposed reference from the i-th word in the set S and the j-th stimuli in the set R, but it is stated that only an association between them exists. Words create activation in different areas in the brain (Pulvermüller, 2003). Nouns tend to activate visual areas, verbs usually activate motor areas if the action can be performed by the individual and visual areas otherwise. The activation of different areas is a result of the associations created with different types of stimuli experienced with the word. The word *write*, for example, is associated with the motor stimuli of the action of writing and the visual stimuli of the objects used in writing. The definition of the meaning of a word is still an open problem, and the complexity in the definition arises from the interaction between different stimuli. It is reasonable to assume that a factor that influences a word frequency of usage is the number of associations with different stimuli. The more stimuli a word is associated to, the highest is its probability of occurrence. The fact that some words have no apparent meaning, such as prepositions, conjunctions and articles, does not imply that they don't have associations with stimuli. Words that apparently don't carry meaning are the words with the highest frequencies, for example, 'the', 'of', 'of',

'to' and 'a'. These words are going to present the more connection with stimuli, which are merely associative and referential links.

The probability that a signal (word) s_i is associated with a stimulus (meaning) r_j, in a given communication system, is given by

$$p(s_i, r_j) = \frac{a_{ij}}{\|A\|} \tag{7.107}$$

where $\|A\|$ is the normalization factor

$$\|A\| = \sum_i \sum_j a_{ij} \ . \tag{7.108}$$

The number of stimuli associated to a given signal s_i is defined as $\mu_i = \sum_{k=1}^{m} a_{ik}$. The probability of the signal s_i is $p(s_i) = \sum_{j=1}^{m} p(s_i, r_j)$. Using this last definition and substituting Equation 7.107 into it, we get

$$p(s_i) = \frac{\mu_i}{\|A\|} \ . \tag{7.109}$$

Similarly, we define the number of signals associated to a certain stimulus by $\omega_j = \sum_{k=1}^{m} a_{kj}$ and the probability of a given stimulus is given by $p(r_j) = \sum_{k=1}^{m} p(s_k, r_j)$ and using the same substitution with Equation 7.107 we get

$$p(r_j) = \frac{\omega_j}{\|A\|} \ . \tag{7.110}$$

What Equation 7.109 states is that a word is used with a probability proportional to the number of stimuli it is associated to. This assumption is supported by the fact that word frequency and number of meanings are positively correlated (Manning and Schütze, 1999). This probability is dependent on the internal organization of the communication system, on the other hand, the Equation 7.110 is related to the frequency of what we talk about, which is dedicated to the outside world. The previous one seems really important in the language communication model, but communication in human language is ofter detached from the here and now (Hockett, 1960). "It is hard to establish from the state of the art of cognitive science whether displaced speech acts are entirely controlled by the internal structure of the communication system or not" (Ferrer-i-Cancho, 2006).

The signal entropy is given by

$$H(S) = -\sum_{i=1}^{n} p(s_i) \log p(s_i) \tag{7.111}$$

and the mutual information between the signal and stimulus is expressed by

$$I(S,R) = \sum_{\substack{i=1 \\ j=1}}^{n,m} p(s_i, r_j) \frac{\log p(s_i, r_j)}{p(s_i)p(r_j)} \ . \tag{7.112}$$

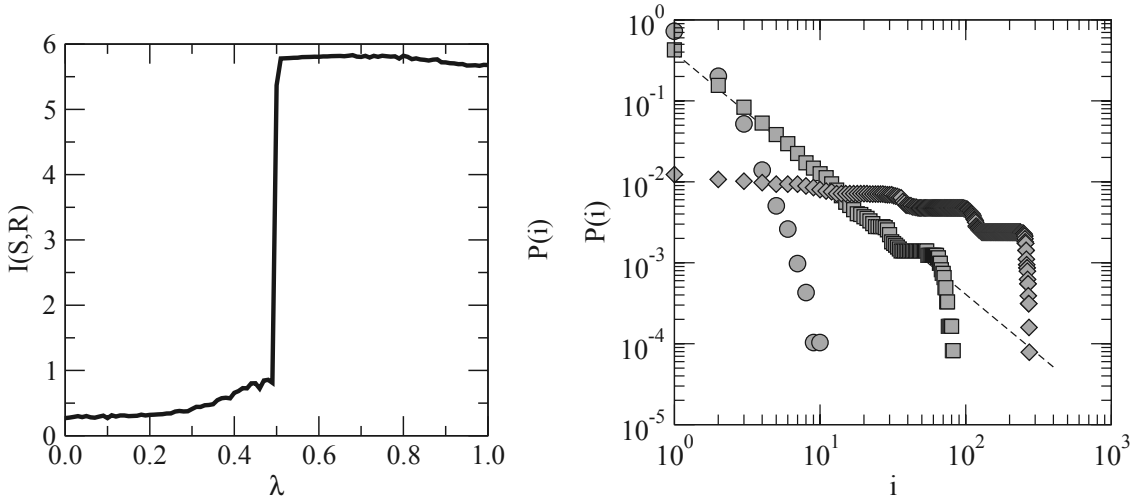
Ferrer-i-Cancho (2005a) proposes to define a function Ω that a communication system must minimize. "The function is a combination of the goal of communication, that is, maximizing the information transfer between the set of signals and the set of stimuli, $I(S,R)$, and the constrains imposed by the biology of the communication system, which tend to minimize $H(S)$, the entropy associated to signals. $H(S)$ is the cost of the communication" (Ferrer-i-Cancho, 2005a). The function Ω is defined as a linear combination of the factors involved,

$$\Omega(\lambda) = -\lambda I(S,R) + (1-\lambda)H(S) \ , \tag{7.113}$$

where the parameter λ is such that $0 \leq \lambda \leq 1$ and controls the balance between the information transference and cost. When $\lambda < 1/2$, the cost of word use is prioritized over the cost of information transfer. The extreme values $\lambda = 0$ and $\lambda = 1$ happen when just one factor is considered, word use and information transfer, respectively.

In order to minimize $\Omega(\lambda)$, Ferrer-i-Cancho (2005a) uses a Monte Carlo algorithm. The results for different values of λ (compromise between word use and information transfer) are presented in the Figure 7.29a. It shows an abrupt jump in the information transfer which takes place at a critical value of $\lambda = \lambda^* = 1/2 - \epsilon$, where ϵ is a small positive value ($\epsilon \approx 0.002$ in Figure 7.29a). The frequency versus rank relation is presented in Figure 7.29b. We observe that Zipf's law is found at the sharp increase in $I(S,r)$ at $\lambda \approx 1/$.

From the results presented in Figure 7.29 and comparing with the frequency-rank relations for phones, diphons, triphones, syllables and words presented Figure 7.3 and 7.1, we conclude that the smaller unities of speech behave as a system whose parameter $\lambda < \lambda^*$, a system in which the cost of use of symbols is prioritized over the cost of information transfer. That result would be expected, since phones, diphones, triphones and syllables don't carry any meaning, or are quite inefficient in matters concerning transmission of information. Among the examples presented here, only a system that use words as its symbols happen to have a balance between signal use and information transfer, and for that reason the Zipf's law is observed in frequency-rank relation on words. We might conclude that words hold a very special position on the language structuring, being responsible to balance information transfer and use.



(a) $I(S,R)$, the information transfer between words and meanings, versus λ, the parameter regulating the balance between maximizing $I(S,R)$ and minimizing the entropy of words.

(b) $P(i)$, the probability of the i-th most likely word in the system for $\lambda = 0.49$ (circles), $\lambda = 0.498$ (squares) and $\lambda = 0.5$ (diamonds). The dashed line contains the theoretical curve for $\lambda = 0.498$.

Figure 7.29: Some computational results on the model where meaning probabilities are governed by the internal structure of the communication system. The size of the system is $n = m = 400$ (i.e. 400 words and meanings). Figure reproduced from Ferrer-i-Cancho (2005a).

7.9 Coherence in Zipf

Cristelli et al. (2012) proposes the idea that a Zipfian distribution is observed in group depending on the number of events and also "requires a fundamental property of the sample distribution which we call 'coherence' and it corresponds to a 'screening' between various elements of the set". They present the classical example of city sizes, where the data from each European country alone has a Zipf law in it, but the data from all European country combined does not produce this power law pattern anymore. Another interesting example present is the evolution of the gross domestic product (GDP) of the countries of the world. The data from 1900 to 2008 shows a progressive approach to a Zipf law, what is hypothesized to be caused by globalization, creating a coherence among world's economies.

A question arises, whether a subset of the original Zipfan distributed set would also present the same power law pattern. The answer to this question is not straightforward, and is seems reasonable to assume that different types cause different contributions to the structure of the group. It is intuitive that a subset of the original set, made of the group of most frequent types, will also present the same power law pattern, since the ratio of two subsequent types will still be the same. Consider the Zipf's equation, rewritten here

for convenience,

$$f(k;s,N) = \frac{1/k^s}{\sum_{n=1}^{N} 1/n^s} = C/k^s \ . \tag{7.114}$$

Discarding the M least frequent types will only produce a change in the constant C, and therefore the relation among the remaining types will be preserved.

Consider now the situation where the most frequent types are withdrawn. This will lead to a change in the relationship among the remaining items. Let's suppose the first k^* most frequent items are removed. We rewrite the Zipf's equation using a new rank variable $k' = k - k^*$. As we remove the most frequent items, the total number of tokens in out dataset will be reduced to N' and we want now to attest what is the new relation between rank and frequency of occurrence, if their product is roughly a constant, then we will find again a Zipf law. Since the dataset, regarding the maintained items, has not changed, their frequencies of occurrence have also not changed. The product between rank and frequency is given by

$$\begin{aligned} k'^s f(k;s,N) &= (k-k^*)^s f(k;s,N) \\ &= k^s \left(1 - \frac{k^*}{k}\right)^s f(k;s,N) \\ &= \left(1 - \frac{k^*}{k}\right)^s C \ . \end{aligned} \tag{7.115}$$

In order to investigate whether that value is a constant or not, we need to evaluate $(1-k^*/k)^s$. As we are considering the case when $k > k^*$, we will have $0 < (1-k^*/k)^s < 1$. If $k >> k^*$, we will have $(1-k^*/k)^s \approx 1$, and the relation between rank and frequency will be approximately a constant. The smallest k we are considering is $k = k^* + 1$, in which case, we shall have $(1-k^*/k)^s = (k^*+1)^{-s}$, what will lead to the strongest deviation from Zipf law. These remarks might be observed in Figure 7.30, where we present the effect of drawing the k^* most frequent types for $k^* = 1, 10, 100, \ldots$. The value of $(1 - k^*/k)^s$ as a function of k is presented in Figure 7.31 for some values of s and k^*.

The new relation between rank and frequency, as we remove the k^* most frequent types is displayed in Figure 7.30 for different values of $k*$. The Zipf law doesn't hold anymore as the most frequent types are removed. In fact, it is enough to remove the most frequent type, to make the Zipf law not valid anymore. This is known as the New York's effect, "we cannot draw two or more 'New York's', for we would destroy the coherence of the set if we did" (Cristelli et al., 2012).

Another possibility is to remove types from within the Zipf plot, and we want to investigate what will be the results on the new created data, whether it will hold a power law relation or not. As previously argued, the highest frequency types, will suffer no

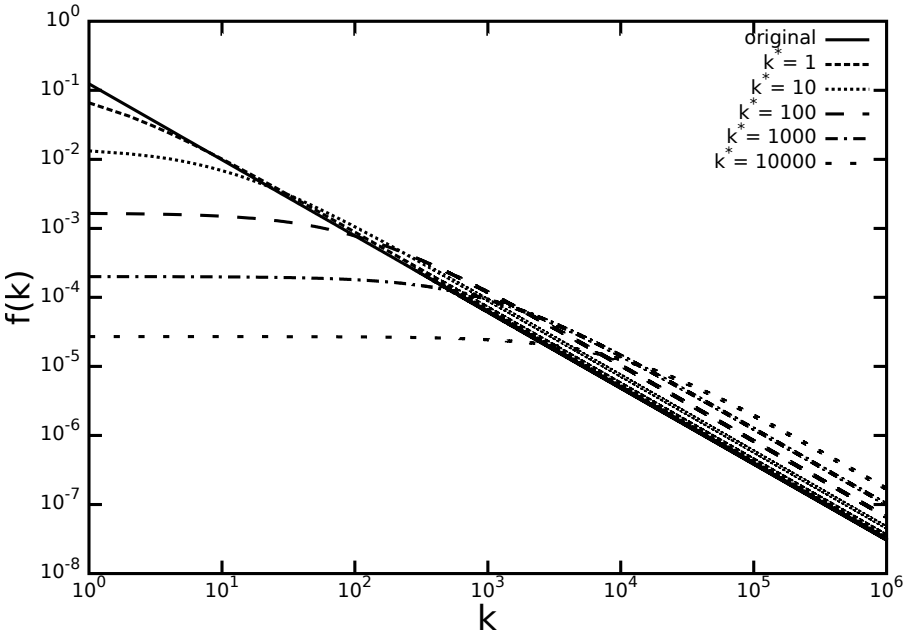


Figure 7.30: New relation between rank and frequency as the k^* most frequent types are removed. Results obtained using an exponent $s = 1.1$.

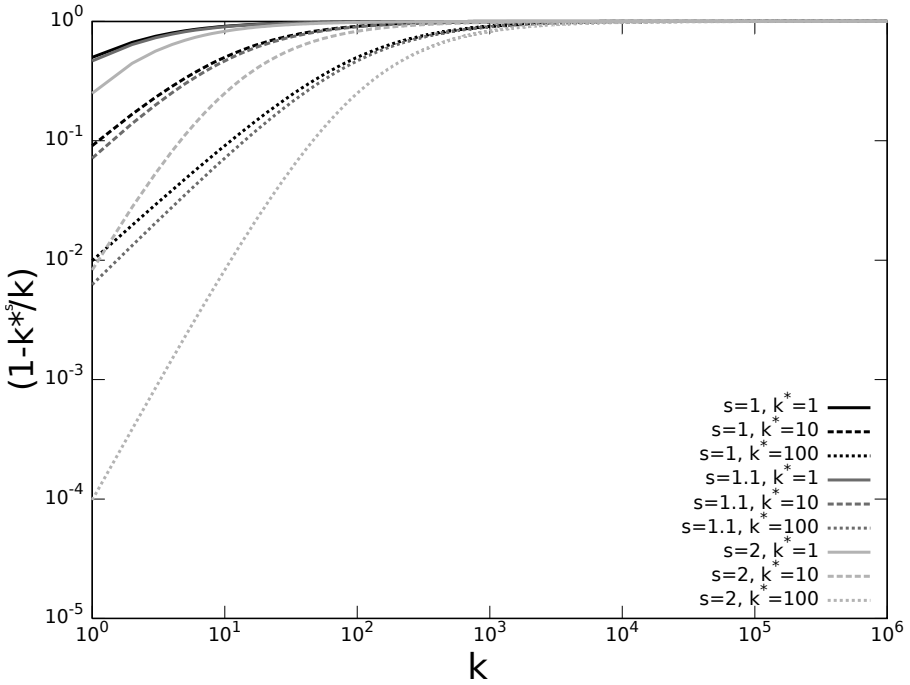


Figure 7.31: Distortion factor as a result of withdrawing the k^* most frequent types.

distortion on the relation they have among them, holding the same Zipf law. The types less frequent than the removed data will suffer a distortion. Consider the removal of some middle frequency types, as explained in figure 7.32. The types with rank between k^* and k^{**} will be removed and the data originally ranked above k^{**} will have a new frequency-rank relation, as might be expressed by

$$\begin{aligned} k'^s f(k;s,N) &= (k - k^{**} + k^*)^s f(k;s,N) \\ &= (k - \Delta k^*)^s f(k;s,N) \\ &= \left(1 - \frac{\Delta k^*}{k}\right)^s k^s f(k;s,N) \\ &= \left(1 - \frac{\Delta k^*}{k}\right)^s C \end{aligned} \tag{7.116}$$

From this relation we might observe that types with $k >> \Delta k^*$ will suffer a small distortion on the relation between frequency and rank. The distortion will be the same as the previously considered in Figure 7.31, but changing k^* by Δk^* in the distortion factor equation.

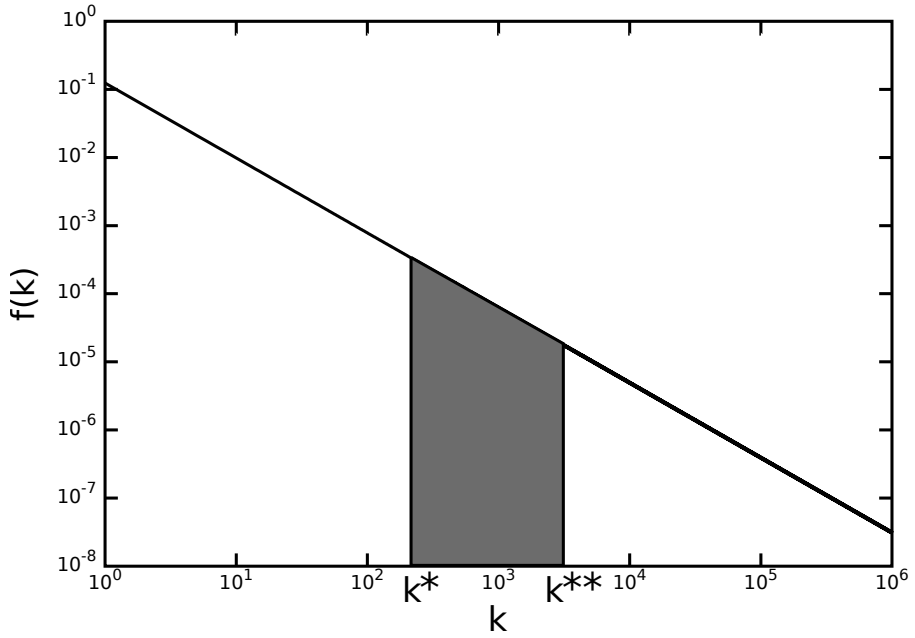


Figure 7.32: Removal procedure of middle frequency types: the types ranked between k^* and k^{**} (gray area) are removed.

Cristelli et al. (2012) suggest that the aggregation of different data, that present alone a Zipf law, will not present the same frequency-rank behaviour, since these data come from different distributions and might not have coherence among them. In order to verify this supposition we have generated random texts using different symbol probabilities.

Each one of them present a Zipf like behaviour, in accordance with Miller (1957) and Li (1992). The random texts created have 6 symbols and one word-break symbol. Symbols were randomly drawn, according to a different set of probabilities for each text, until the text length reached 8×10^4 symbols. Figure 7.33 present the Zipf relation observed in each random generated data and also in the large data created by the concatenation of the previous seven, which also presents a Zipf law. This could contradict what was expected by Cristelli et al. (2012) or could corroborate that random text don't present a truly Zipfian distribution pattern, the exponential decreasing frequency is just again a mere byproduct of selecting rank as the independent variable.

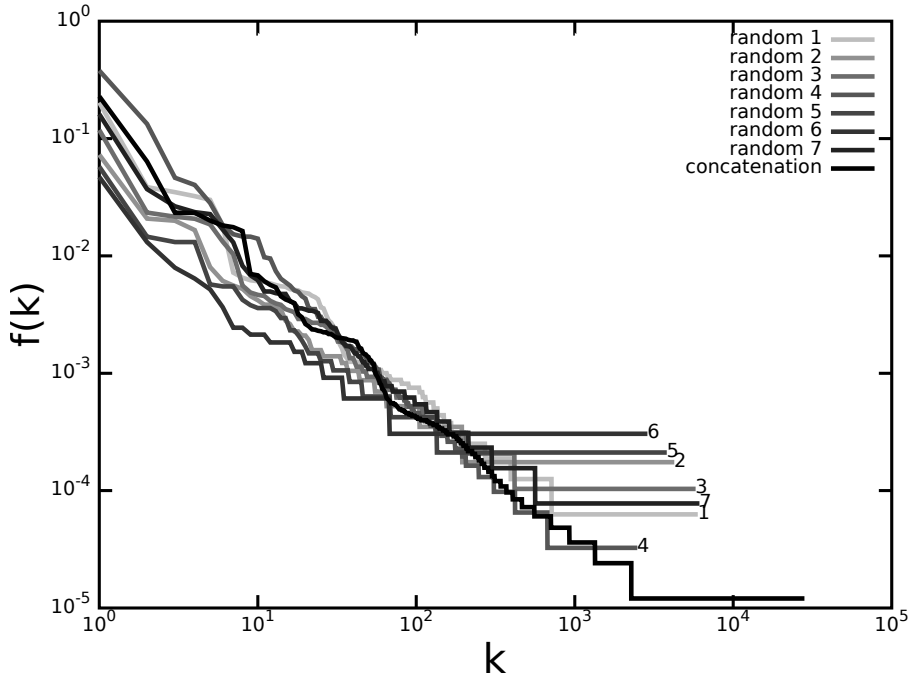


Figure 7.33: Frequency-rank relation of seven random generated texts, with different symbols probabilities, and the frequency-rank relation in the concatenation of those seven random texts.

7.10 Generalized Zipf

The frequency rank relation suffer observable deviations from the power-law on the high- and low-rank regions. The former one is usually explained as an effect of under-sampling of rare words in the corpus, what create the stair case pattern. The last one is observed as a flattering in the upper part of the Zipf curve. To account for that behavior on the low-rank region, Mandelbrot (1965) proposed a modified formula, given a generalized relation between rank and frequency of occurrence

$$f(k; s, N, q) = C(k + q)^{-s} \ , \tag{7.117}$$

where the new parameter q also characterize the distribution at hand and the constant C is now given by

$$C = \frac{1}{\sum_{n=1}^{N}(n+q)^{-s}} \ . \tag{7.118}$$

Note that this generalization becomes the previous Zipf's Law when $q = 0$. Asymptotically, the relation 7.117, also known as Zipf-Mandelbrot Formula, is equivalent to the power law 7.9, when $k >> q$. The constant C is the inverse of the generalized harmonic number, which has a limit as N tends to infinity only if $s > 1$. It might also be called as the Riemann zeta function, when $N \to \infty$: $\zeta(s) = \sum_{n=1}^{\infty} n^{-s}$. When $s = 1$ we have the harmonic series: $\zeta(1) = 1 + \frac{1}{2} + \frac{1}{3} + \cdots = \infty$; when $s = 2$ the we get $\zeta(2) = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots = \frac{\pi^2}{6}$, which demonstration is known as the *Basel problem*, first posed by Pietro Mengoli in 1644 and solved by Leonhard Euler in 1735. As our constant C tends to the inverse of the Riemann zeta function in the limit, it might be also seen as Dirichlet series over the Möbius function $\mu(n)$, that means

$$\lim_{N\to\infty} C = \frac{1}{\zeta(s)} = \sum_{n=1}^{\infty} \frac{\mu(n)}{n^s} \ . \tag{7.119}$$

Miller (1957); Mandelbrot (1965) propose the model of *intermittent silence* to study the statistical properties of languages. Mandelbrot proposes that the cost of producing a word is proportional to its length in characters, and defined the information content of a word as the Shannon entropy. Minimizing the average cost per unit of information, he showed that the random placement of spaces leads to Zipf's rule, which is actually the optimal solution, which might be achieved as a result of language evolution. The use of *space* as a word-delimiter would cause an exponential increase in the number of words regarding its length, and that would hold even if the characters used don't are not equiprobable (Li, 1992). That behavior, on the contrary, is not observed, and the relation between the number of existing words and their length is not even monotonically increasing, as we observed in the results presented in Figure 7.7.

The generalization proposed in 7.117, is more suited to follow the empirical observations. According to Mandelbrot (1965), the flatness presented in the low-rank regions depends on the 'wealth of vocabulary' of the subject, and the parameters C, q and s are dependent on the subject, and they "do not characterize a language, although it may well be that different languages favor different ranges of values for the parameters".

Manin (2009) present a minimization procedure of the cost ratio $C^* = C/H$, where C is the average cost of producing words and H is the entropy per word. Manin (2009) shows that there is one Ansatz that leads to the Zipf's law or the Zipf-Mandelbrot's law. If we consider that each word w_k happens with probability p_k, we assign an unspecified

value C_k to the cost of producing the word w_k. The word's information content is given by $H_k = -\log_2 p_k$. The average cost per word is then calculated by

$$C = \sum_k p_k C_k \tag{7.120}$$

and the average entropy per word by

$$H = -\sum_k p_k \log_2 p_k \ . \tag{7.121}$$

The probabilities of the words are such that $\sum_k p_k = 1$ must hold. We wish to minimize the cost ratio $C^* = C/H$, so the Lagrange multipliers are used, defining the Lagrange function

$$\Lambda(p_k, \lambda) = C/H + \lambda \cdot \left(\sum_k p_k - 1 \right) \ . \tag{7.122}$$

Finding the minimum of the cost ration is equivalent to find the stationary point for the Lagrange function, which correspond to those points where the partial derivatives of Λ are zero, i.e. $\nabla\Lambda = 0$. That means we seek for

$$\frac{\partial}{\partial p_k} \left(C^* + \lambda \sum_j p_j \right) = 0 \ . \tag{7.123}$$

Performing the differentiation, we obtain

$$\frac{C_k}{H} + \frac{C}{H^2} \left(\log_2 p_k + \frac{1}{\ln 2} \right) + \lambda = 0 \quad , \forall k \ . \tag{7.124}$$

The frequency p_k is given by

$$p_k = 2^{-\lambda H^2/C} 2^{-1/\ln_2} 2^{-C_k H/C} \ . \tag{7.125}$$

A power law for frequencies could only result from the Ansatz

$$C_k = C_0 \log_2 k \ , \tag{7.126}$$

what leads to

$$p_k = 2^{-\lambda H^2/C} 2^{-1/\ln_2} k^{-C_0 H/C} \ , \tag{7.127}$$

that might be rewritten in the Zipf form by $p_k = C' k^{-s}$, where the constant are defined by $C' = 2^{-\lambda H^2/C} 2^{-1/\ln_2}$ and $s = C_0 H/C$.

According to Manin (2009), the relation 7.126 "is a much more plausible argument

(...) which does not depend on any assumptions about word length at all. Suppose words are stored in some kind of addressable memory. For simplicity, one can imagine a linear array of memory cells, each containing one word. Then, the cost of retrieving the word in the kth cell can be assumed to be proportional to the length of its address, that is to the minimum number of bits (or neuron firings, say) needed to specify the address. And this is precisely $\log_2 k$"(Manin, 2009).

Using a different Ansatz

$$C_k = C_0 \log_2(k+q) \ , \tag{7.128}$$

that would lead to the Zipf-Mandelbrot's law. This law is "obtained from a model optimizing the information/cost ratio with no assumptions about word lengths. This model is not equivalent to the random typing model, and allows the optimum to be achieved via local dynamics, i.e. in a causal, rather than teleological manner. However, the two parameters of the resulting distribution, s and q, are not independent, and as a result, it does not provide a reasonable fit to the empirical data"(Manin, 2009).

7.11 Heaps' Law

Zipf's law is perhaps the best evidence of a universal physical law, which has drawn a considerable attention on the academia. Many explanations of Zipf's law has been given: as a result of a random process (Miller, 1957; Li, 1992), or due to the principle of least effort (Zipf, 1949; Ferrer-i-Cancho and Solé, 2003), or a Boltzmann-type approach (Düring et al., 2008), or avalanche dynamics in a critical system (Bak, 1999).

Power laws are referred to have a scale invariance behaviour, making it impossible to define a characteristic scale. Language statistics present more that one dimension, we might consider the frequency of occurrence as a independent variable, or the text (or utterance) length T (the number of tokens in a sample), or even the vocabulary size V_T (the number of types present in a sample). The relation between text length T and vocabulary size V_T is known as Heap's law (also called Herdan's law), which states that the vocabulary V_L grows in a sublinear form with the text length T,

$$V_T \propto T^{\alpha} \ , \quad \alpha < 1 \ . \tag{7.129}$$

Leijenhorst et al. (2005) shows that Heaps' law and the generalized Zipf's law are related. Assuming a Mandelbrot distribution, it is possible to mathematically derive Heaps' law. If both Zipf's law and Heaps' law hold, than their exponents are related by $\alpha = 1/s$. It means that the Zipf's exponential coefficient must be greater then one, in order to make $\alpha < 1$, a plausible value. As we observed previously, the Zipf exponent for

natural phenomena always presented an exponent greater than one, what differed from the random generated texts. In fact, Lü et al. (2010) shows that the relation $\alpha = 1/s$ is only an asymptotic solution that every large-size system holds when $s > 1$.

A formal derivation of Heaps' law is given by Leijenhorst et al. (2005) and we present it here. For that purpose we are going to call $\mathcal{W}$ the set of words in a text (the set of tokens in a sample) and $\mathcal{D}$ is the set of different words in a text (the set of types in a sample). For simplicity reason, we shall consider that the occurrence of each word is independent one from another. We are characterizing a text as a random process of independently drawing words from a set with replacement. After n words were drawn, we will end up with a set $\mathcal{W}_n$ with length n and a set $\mathcal{D}_n$ with length $m_n \leq n$. The Heaps' law states a relation between m_n and n, and we are particularly interested in the behaviour as $n \to \infty$. In that case, we shall deal with $\mathcal{D}$ as the underlying lexicon of a language, the set of all possible types. In order to know what is the expected number of different words in sequence of independently tokens drawn from a set, we shall analyze the n-th draw in detail.

As the n-th token is drawn, there might be two possibilities: 1) the n-th token is a new type, and therefore is not present in $\mathcal{D}_{n-1}$; or 2) the n-th token is already in $\mathcal{D}_{n-1}$, and therefore the length of $\mathcal{D}_n$ is equal to the length of $\mathcal{D}_{n-1}$. After n draws were made, we shall consider w_n as the n-th token in the sequence, and we want to find what is the probability that there are a different types, $\Pr(m_n = a)$. There may not be more types than tokens, so $\Pr(m_n = a) = 0$ if $a > n$. If $a \leq n$ we shall have

$$\begin{aligned}
\Pr(m_n = a) &= \Pr(m_{n-1} = a - 1 \wedge w_n \notin \mathcal{D}_{n-1}) + \Pr(m_{n-1} = a \wedge w_n \in \mathcal{D}_{n-1}) \\
&= \Pr(m_{n-1} = a - 1) \Pr(w_n \notin \mathcal{D}_{n-1}) + \\
&\quad \Pr(m_{n-1} = a) \Pr(w_n \in \mathcal{D}_{n-1})
\end{aligned} \tag{7.130}$$

We know that $\Pr(w_n \notin \mathcal{D}_{n-1}) = 1 - \Pr(w_n \in \mathcal{D}_{n-1})$, and the former one might be given by the sum of the probabilities of every word in the underlying lexicon not belonging to the set $\mathcal{D}_{n-1}$

$$\begin{aligned}
\Pr(w_n \notin \mathcal{D}_{n-1}) &= \sum_{i \in \mathcal{D}} \Pr(w_n = i \wedge i \notin \mathcal{D}_{n-1}) \\
&= \sum_{i \in \mathcal{D}} \Pr(w_n = i) \Pr(i \notin \mathcal{D}_{n-1}) \\
&= \sum_{i \in \mathcal{D}} p_i (1 - p_i)^{n-1}
\end{aligned} \tag{7.131}$$

where p_i stands for the underlying probability of the i-th word in the lexicon, and $\Pr(i \notin \mathcal{D}_{n-1})$ is the probability that the i-th word has not happened after $n-1$ draws, so it must be equals to $(1 - p_i)^{n-1}$.

For convenience, We are going to use the same notation as Leijenhorst et al. (2005)

$$S_n = \sum_{i \in \mathcal{D}} p_i (1 - p_i)^{n-1} \tag{7.132}$$

$$M_n = \sum_{i \in \mathcal{D}} (1 - p_i)^n \tag{7.133}$$

from which we might find the relation bellow

$$\begin{aligned} M_{n-1} - M_n &= \sum_{i \in \mathcal{D}} (1 - p_i)^{n-1} - \sum_{i \in \mathcal{D}} (1 - p_i)^n \\ &= \sum_{i \in \mathcal{D}} (1 - p_i)^{n-1} \left(1 - (1 - p_i)\right) \\ &= \sum_{i \in \mathcal{D}} p_i (1 - p_i)^{n-1} = S_n \ . \end{aligned} \tag{7.134}$$

Using $N(n, a) = \Pr(m_n = a)$ we have $N(1, 1)$, since only one token was drawn. As previously stated, $N(n, a) = 0$ if $n < a$ and for $n \geq a$ shall have

$$N(n, a) = N(n - 1, a - 1) S_n + N(n - 1, a)(1 - S_n) \tag{7.135}$$

what suggests a recurrence relation to find the value of $N(n, a)$.

After n tokens were drawn, the expected number of types is given by

$$\begin{aligned} E[m_n] &= \sum_{a=1}^{n} a N(n, a) \\ &= \sum_{a=1}^{n} a \left(N(n - 1, a - 1) S_n + N(n - 1, a)(1 - S_n) \right) \\ &= S_n \sum_{a=1}^{n} a N(n - 1, a - 1) + (1 - S_n) \sum_{a=1}^{n} a N(n - 1, a) \\ &= S_n \sum_{a=1}^{n} a N(n - 1, a - 1) + (1 - S_n) \left[n N(n - 1, n) + \sum_{a=1}^{n-1} a N(n - 1, a) \right] \\ &= S_n \sum_{a=1}^{n} a N(n - 1, a - 1) + (1 - S_n) \left[0 + E[m_{n-1}] \right] \\ &= S_n \left[\sum_{a=1}^{n} (a - 1) N(n - 1, a - 1) + \sum_{a=1}^{n} N(n - 1, a - 1) \right] + (1 - S_n) E[m_{n-1}] \\ &= S_n E[m_{n-1}] + S_n \sum_{a=1}^{n} N(n - 1, a - 1) + (1 - S_n) E[m_{n-1}] \end{aligned}$$

$$= S_n \sum_{a=1}^{n} N(n-1, a-1) + E[m_{n-1}]$$
$$= S_n + E[m_{n-1}] \tag{7.136}$$

Leijenhorst et al. (2005) presents a formal proof that when the data is Zipfianly distributed, it will lead to a Heaps behaviour on the increase number of types as the number of tokens increase. We might wonder weather this sort of behaviour is characteristic only of Zipfian distributed data, or whether it could happen with other type of distributions, or even with every type of distribution. We present in the Figure 7.34 a computer simulation of the expected number of types as the number of tokens increases, using three different discrete probability distributions: binomial, uniform and Zipf. For every one of them we considered a sample of length 10^6 and a lexicon of size 10^4.

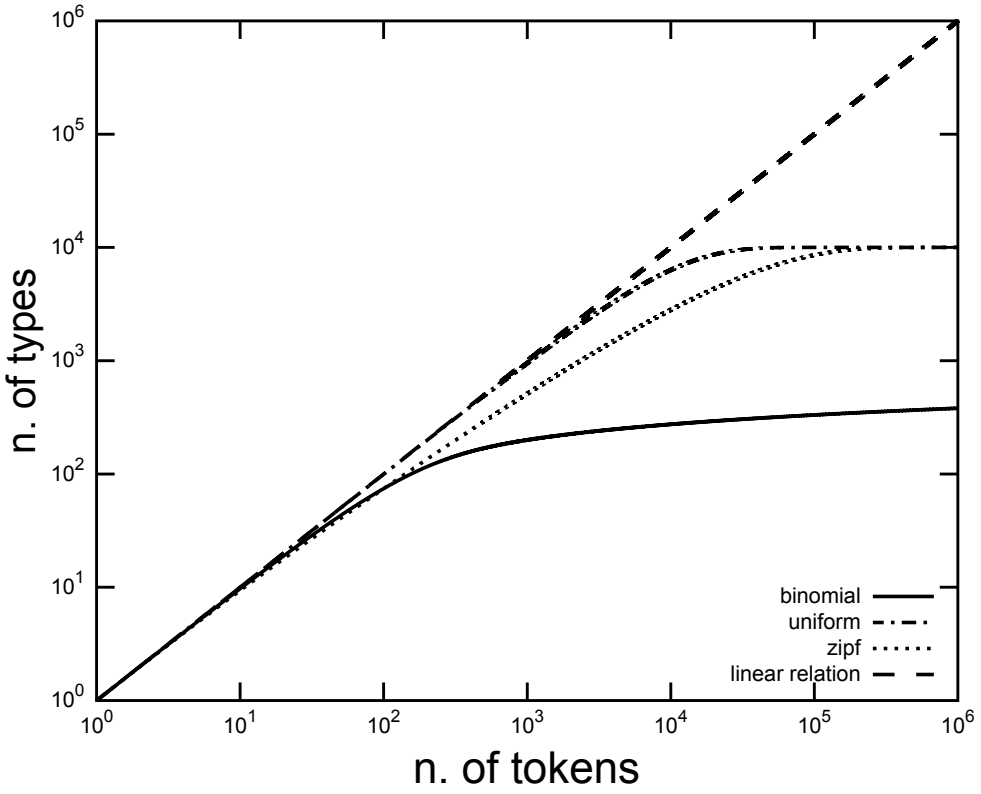


Figure 7.34: The recurrence equation is used to estimate the expected number of types for a sample with a certain number of tokens. Different probabilities for set were used: binomial, uniform and Zipf. All of them present a Heaps like behaviour.

This behaviour is expected, since the sequence of S_n used to compute $E[m_n]$ is such that

$$1 > S_1 > S_2 > \cdots > S_{n-1} > S_n > 0 \ . \tag{7.137}$$

What might be verified by taking S_n and writting it as a function of S_{n-1}.

$$\begin{aligned} S_n &= \sum_{i=1}^{N} p_i(1-p_i)^n \\ &= \sum_{i=1}^{N} p_i(1-p_i)^{n-1}(1-p_i) \\ &= \sum_{i=1}^{N} p_i(1-p_i)^{n-1} - \sum_{i=1}^{N} p_i^2(1-p_i)^{n-1} \\ &= S_{n-1} - \sum_{i=1}^{N} p_i^2(1-p_i)^{n-1} = S_{n-1} - R \end{aligned} \tag{7.138}$$

The latest term R is a sum of positive values, and therefore is positive itself. We may conclude then that $S_n > S_{n-1}$, and the sequence of S_n's is a monotonically decreasing sequence with increasing n.

Using Equation 7.136 and the fact that $E[m_0] = 1$ and $S_0 = 1$ we might write $E[m_n]$ in the following way:

$$E[m_n] = \sum_{i=0}^{n} S_i \ . \tag{7.139}$$

The values of S_i's are finite and therefore, for a finite sample, the number of types observed is also finite. Considering the assyntotic behaviour, as $n \to \infty$, we want to find the underlying number of types (lexemes) in a language vocabulary (lexicon). To verify if the sum in Equation 7.139 converges, me might use the D'Alembert's criterion, which states that

$$\lim_{n\to\infty} \left| \frac{S_n}{S_{n-1}} \right| = r \tag{7.140}$$

and the sum will converge if $r < 1$, it will diverge if $r > 1$ and it will be unconclusive if $r = 1$.

We might then use the relation expressed in Equation 7.138 and express Equation 7.140 as

$$\begin{aligned} \lim_{n\to\infty} \left| \frac{S_n}{S_{n-1}} \right| &= \lim_{n\to\infty} \left| \frac{S_{n-1} - R}{S_{n-1}} \right| \\ &= \lim_{n\to\infty} \left| 1 - \frac{R}{S_{n-1}} \right| = r \end{aligned} \tag{7.141}$$

As $S_{n-1} > R > 0$, we may conclude that the sum will converge.

We conclude then that the behaviour of the expected number of types in a sample is a monotonically increasing function of the sample size, converging to the underlying number of types in the vocabulary in question. Although the lexicon size may increase

as the text length increases infinitely, the lexicon size has a limiting value. The expected value of the lexicon size is still finite and may be estimated given the distribution is known. The lexicon growth behaviour in natural language is verified for some texts from the Guternberg Database. Figure 7.35 bellows present the relation between the text length and the vocabulary size for 35 different books, all plotted in gray. The continuous black curve presents the relation for the metabook created by the concatenation of all 35 books. As a reference, it is also presented, by a dashed line, the 1:1 relation.

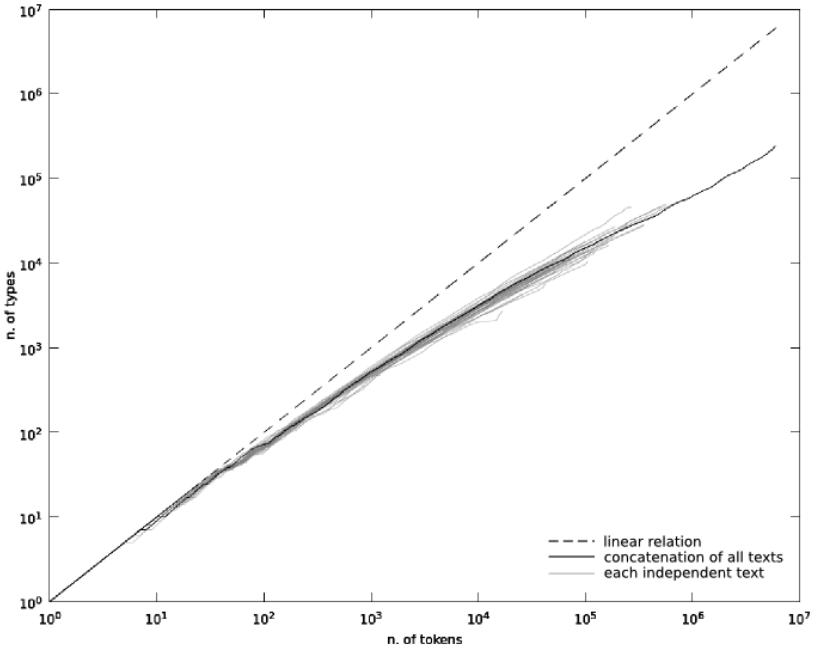


Figure 7.35: The relation between the number of tokens and types in 35 books from Gutenberg Database is presented in gray. The dark curve is the result when all 35 books are concatenated and the dashed line is presented only as a reference, when the number of types is equals the number of tokens.

8

Feature Theory

> Differences which have differentiating value are, as we have seen, more accessible to perception and to memory than differences which have no value at all, but on the other hand differences between phonemes—since they lack particular meanings—strain perception and memory and necessarily require a great deal of them. We would expect, therefore, that the number of these primordial and unmotivated values would be relatively small for any given language.
>
> Roman Jakobson (1942)

Speech sounds are continuous signals in the physical world, but somewhere on the communication process they are interpreted by humans as discrete categories. This process of perceiving physical world stimuli into a discrete set of categories is called categorical perception (CP). It might be regarded as a sensory phenomenon of percept invariance when experiencing a stimulus that is varied along a continuum. According to the CP

view a phenomenon tend to be perceived in terms of the categories that we have formed. Multiple stimuli are mapped into a single category and there is a sharp change of perception at a certain point in the continuum along which a stimulus may change (Liberman et al., 1957). Our perceptions are biased such that differences between objects that belong to different categories are accentuated, and differences between objects that fall into the same category receives little attention. That means, humans have a great ability to perform interclass segregation, but very poor capability to segregate intraclass stimuli. CP has been found in a wide range of stimuli, including the perception of color boundaries (Bornstein and Korda, 1984), phonemes (Liberman et al., 1957) and facial expressions (Etcoff and Magee, 1992). In speech, categorical perception suggests that speech perception involves a phonemic encoding step, where the perceptual input is represented in terms of discrete phonemic category labels Liberman et al. (1957). According to Iverson and Kuhl (2000) the encoding process could distort perception stretching distances in regions in which stimuli are equidistant between prototypes (i.e., at phonemic boundaries) and shrinking distances near prototypes.

CP is an important mechanism for processing of naturally occurring stimuli in a continua. It is important to perceive and respond to objects and events in an environment, in certain situations, it might be critical to the survival. The ability to ignore redundant sensory variation and cope with the relevant information is necessary to reduce the processing demands and deal only with relevant information. The variation of stimuli within a category is much smaller than the variation across different categories. Creating categories and grouping objects is a task that simplifies further processing.

A good illustration of categorization happens when we observe the colors of a rainbow. Although we though it is made of a smooth range of light frequencies, we perceive the rainbow as bands with distinct colors. If we compare the colors in the rainbow we realize it is easier to distinguish two different shades of colors when they come from different color boundaries. When they are from the same color boundaries the task is harder, even if we select shades of colors with the same frequency difference of the previous pair (Bornstein and Korda, 1984).

The phonemes of a language, as proposed by Saussure, are seen as atomic symbols, indivisible by its nature. The idea of indivisible sound units capable of forming meaningful strings was introduced into the Greek philosophical literature under the name *stoicheia*. "The sound shape of language and correspondingly its alphabet were viewed as a joint coherent system with a limited number of discrete and interconnected formal units. This concept proved to be so persuasive that Democritus (fragment A6; cf. Diels and Wilpert) and his adherent Lucretius, in searching for an analogy which might confirm their theory of the atomic structure of the physical universe, cited *stoicheia* as the minimal components of

speech"(Jakobson and Waugh, 2002). One question that rises is whether this belief holds or not in reality. Is it possible to divide and analyze the phonemes using some criteria? Those questions bring along the doubt on what speech sounds are. We know humans can produce many other sounds that are not used in speech, and many sounds, unimaginable to be used as speech sounds in one culture, may turn out to be part of some languages' speech sounds inventory. In many languages of southern Africa, sound clicks are speech sounds used as consonants in the language. The click sounds are obstruents articulated with two closures. There are five places of articulation at which they occur: dental, lateral, bilabial, alveolar and palatal. The IPA symbols used for them are, respectively: [ǀ], [ǁ], [ʘ], [ǃ] and [ǂ]. Some sounds produced by the vocal tract are known not to be used in any language, like whistles, inhalation or a labiolingual trill[1] (a.k.a. "blowing a raspberry"). Those sounds, in other cultures are just regarded as sounds produced with the voice apparatus, but are not used as speech sound at all, although they might be used on a non linguistic communication.

The sounds used in speech communication are inserted into a continuous of infinitesimal variations in its qualities. What we label as a specific symbol or another may be understood as segmentation of this continuous in a small finite set of unities that will be used as the symbols to build speech communication. Consider, for example, the vowels [e] and [ɛ]. What is perceived in each culture may be slightly different, the prototypical [e] in one language may be lower than the prototypical [e] in another language, but not low enough to be considered as a [ɛ] by any of those languages. The place where different cultures imposes a threshold to segregate different vowels in a continuous may differ. Those differences among languages may be perceived when someone is learning a foreign language. Even if the same sounds are shared between one's mother tongue and a certain foreign language, the discretization process to create prototypes in each language may not be the same, and some slight noticeable differences may appear, what makes it hard to speak a foreign language with no accent at all. Peterson and Barney (1952) show that, although there is a great correlate between vowel categorization by listeners and vowels physical properties (frequency of the first and second formants, in this case), the border among them is not sharp. There is a fuzzy transition among all the perceived vowels, and in certain regions it is hard to tell whether the physical observation is of one vowel or another. This is well illustrated in the Figure 8.1.

What is important, under the concern of phonology, is what are speech sound differences capable of being differentiated in a language and eliciting different meaning. Answering this question is the same as answering the question 'what is a possible phoneme?'.

[1]The sound of blowing a raspberry or making a bronx cheer is the sound made by sticking out the tongue between the lips and blowing to make a sound reminiscent of flatulence.

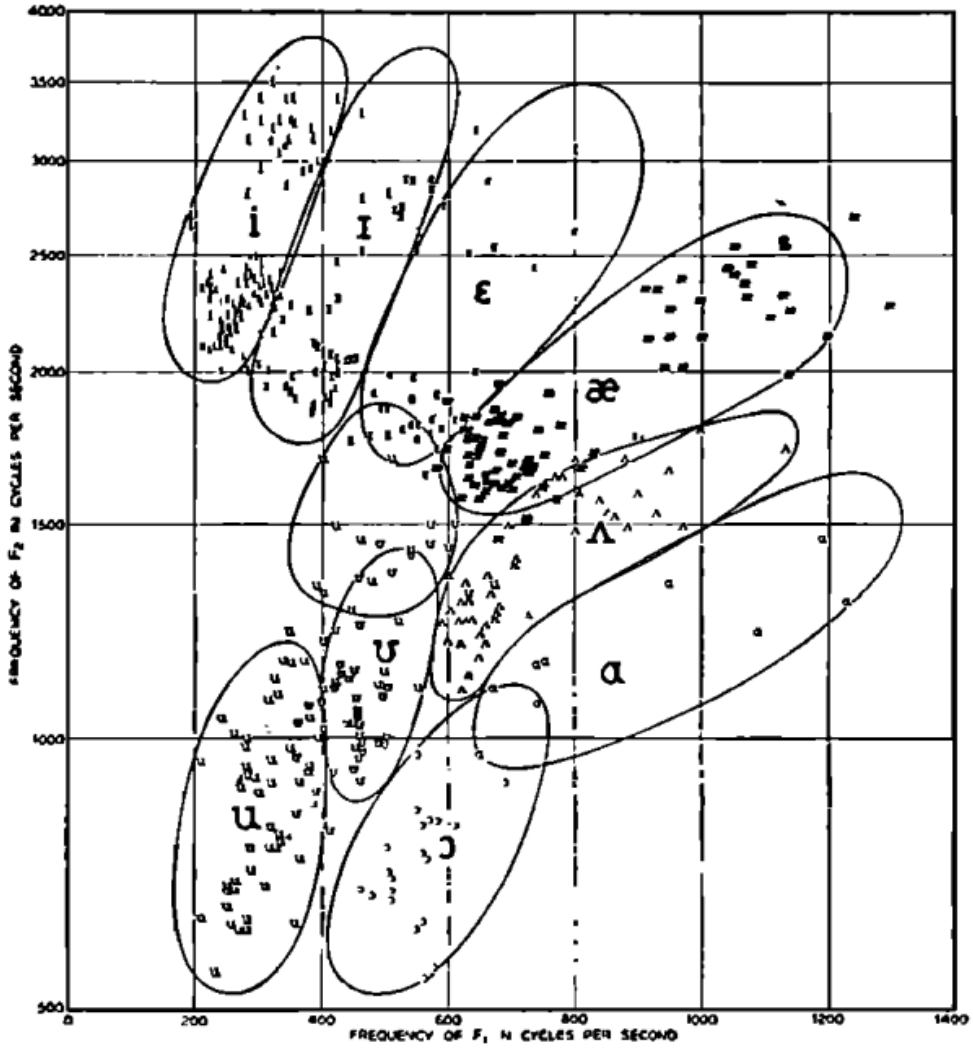


Figure 8.1: Frequency of second formant *versus* frequency of first formant for vowels spoken by men and children, which were classified unanimously by all listeners. (Figure reproduced from (Peterson and Barney, 1952))

The answer to these questions is not trivial, and it involves not only the way humans perceive speech under a certain culture, but also how human perceive the world. Some cognitive aspects of perception are universal. The very fact of perceiving the continuous world as a discrete set of types instead of tokens[2] is a commonplace in speech perception as well as in color perception. If we consider the way humans perceive a rainbow, the categorization into seven bands of colors is an artifact of human perception, a rainbow is in fact a continuous spectrum of colors. That seems like the magical number seven playing its role in human categorization. According to the Miller's Law (Miller, 1956), the number of objects, on average, that a human can hold in working memory is 7 ± 2.

[2]The distinction between *types* and *tokens* is an ontological one, it is the difference between a general sort of things and a concrete instance of this sort. We say then that *tokens* are instances of a *type*.

The way how speech sounds are perceived is rather subjective. One study performed by Laboissiere (2010) shows how the phonemic distinction between /e/ and /ɛ/ takes place for German speakers. A series of stimuli was presented to the subjects of this experiment. The stimulus were a computer generated continuous transition from [e] to [ɛ] that were presented in the context of the minimal pairs in German *fell* ([fɛl]) vs. *fehl* ([feːl]). They were asked to indicate when the transition from one word into another happened. This experiment shows that in the enormous continuum between one prototypical phonemic realization and another, all the instances found are perceived either as one prototype or another, in this wide space there is place only for two possible categorizations.

It is important to note that the contrast shown by Laboissiere (2010) is between two vowels of different quality and duration: [eː] and [ɛ]. In German, the short vowels [i, y, u, e, ø, o] occur only in unstressed syllables of words borrowed from other languages. They are usually considered complementary allophones together with their long counterparts which cannot occur in unstressed syllables. In this situation, it is difficult to say what is the share of contribution of tenseness (tense vs. lax) and length (short vs. long) contrasts into the categorization process.

"The striking thing about phonology is that the infinite phonetic variety in the utterances of any language can be reduced to a small inventory of contrastive units or phonemes"(Hualde, 2004). This inventory is created through a linguist investigative analysis in order to state the building blocks that will assemble one language. This analysis process is many times controversial, it is difficult to establish a consensual phonemic contrast. Even at the phonetic level, it is not easy to choose the right representation, since speech utterances may have different realizations. Hualde (2004) shows an example in Basque where the word *mollako* is pronounced in many ways, regarding the sound of the consonant *k* that is usually voiceless but sometimes voiced and other times partially voiced. That brings a question on which phonetic transcription to choose. This problem is also not solved at the phonemic level because the phonemic status of some palatal consonants in Basque is controversial. The word *mollako* could be then transcribed in two ways: /moʎako/ or /moilako/. "In the relevant Basque dialects /l/ palatalizes after /i/ and palatal glides are absorbed by a following palatal lateral, so that both phonemic inputs would result in the same output"(Hualde, 2004). Just as in other domains, a categorization process may involve different levels. Ladd (2002) shows that exists, in French and Italian, a special link between the mid vowels /e/ and /ɛ/ that is not found between /e/ and /i/, as previously remarked by Trubetzkoy (1939). Category boundaries may be fuzzy and in certain aspects multi-level, allowing overlaps between them. As observed by Hualde (2004), the questionable phonemes in Spanish (j and j̬) followed by an [a] may be mistaken by the hiatus sequence ia. The overlap shown by this example is dependent of

the Spanish dialect, what shows that the categorization process may be done in different ways. As previously remarked by de Saussure (1916), "En matière de langue on s'est toujours contenté d'opérer sur des unités mal définies"[3].

The distinctive feature theory rises as a way to characterize the speech sounds in term of features and use those features to derive phonological explanation of how languages work. It is impractical to establish phonological rules based on physical descriptions of speech sounds. For example, a rule that says that a vowel sound suffer an increase in its first formant frequencies by 205Hz and by 430Hz in its second formant frequency when preceded by another vowel whose first formant frequency is lower than 410Hz, is so much an inefficient way of describing a phonological processes. It is much simpler to describe it in another fashion, in a qualitative form, because we know that the relationships, not the absolute values of physical properties of speech sounds, are important. It is much easier to say that a near-front vowel is transformed into a front vowel when preceded by a front vowel. This qualitative rule is quite simpler and may be used to describe the same phonological rule that is slightly different in distinct languages.

The properties used to describe speech sounds in the distinctive feature theory are general and abstract, that means, their behavior is fuzzy. The theory proposes the existence of a small finite set of features that may be used to analyze speech sounds, in such way that the description of each speech sound in terms of these features is unique. The analysis into distinctive features is an unambiguous process to associate speech sounds into arrays of features that describe these sounds. The classical statement of distinctive features was brought by Chomsky and Halle (1968).

The major class features, proposed by Chomsky and Halle (1968) are: sonorant, vocalic (syllabic) and consonantal. Those major features provide a rough initial grouping of speech sounds into functional types that includes the consonant/vowel distinction. The sonorant sounds are those in which the vocal tract configuration makes spontaneous voicing possible. In syllabic sounds the constriction does not exceed that of high vowels, and spontaneous voicing is possible. They form a syllable peak and stress may be then applied. The consonantal sounds are those produced with a major obstruction in the mid-sagittal region of the vocal tract.

The set of distinctive features proposed by Chomsky and Halle (1968) was afterwards modified by other researchers according to their convenience to better analyze a certain language. Some commonly used features are: consonantal, syllabic, sonorant, continuous, delayed release, nasal, lateral, anterior, coronal, high, back, rounded, low, voiced, tense, strident, ATR (advanced tongue root). The meaning of each feature is described bellow:

syllabic / non-syllabic : The syllabic feature characterizes sounds which have a sonor-

[3]In language's matter it has always been sufficient to operate on ill-defined units.

ity peak within, and we say they constitute a syllable peak. The feature [+syllabic] refers to vowels and syllabic consonants. Contrastive examples of syllabic consonants in English are:

syllabic	non-syllabic
coddling [kɒdl̩ɪŋ]	codling [kɒdlɪŋ]
Hungary [hʌŋgr̩i]	hungry [hʌŋgri]

consonantal / non-consonantal : Consonantal sounds are produced with vocal tract constriction, a radical obstruction in the mid-sagital region. Non-consonantal sounds are produced without such an obstruction.

sonorant / obstruent : "Sonorants are sounds produced with a vocal tract cavity configuration in which spontaneous voicing is possible" (Chomsky and Halle, 1968) due to the air pressure on both sides of any constriction to be approximately equal to the air pressure outside the mouth. Obstruent sounds are characterized by a significantly greater air pressure behind the constriction. [+sonorant] refers to vowels and approximants (glides and semi-vowels); [−sonorant] refers to stops, fricatives and affricates.

coronal / non-coronal : "Coronal sounds are produced by raising the tongue blade toward the teeth or the hard palate; noncoronal sounds are produced without such a gesture" (Chomsky and Halle, 1968). Some authors argue that this feature applies only to consonants; others argue it may apply to front vowels as well. [+coronal] refers to dentals (not including labio-dentals) alveolars, post-alveolars, palato-alveolars, palatals. [−coronal] refers to labials, velars, uvulars, pharyngeals.

anterior / posterior : "Anterior sounds are produced with a primary constriction at or in front of the alveolar ridge. Posterior sounds are produced with a primary constriction behind the alveolar ridge" (Chomsky and Halle, 1968). This feature distinguishes coronal sounds produced in front of the alveolar ridge from those produced behind it. [+anterior] refers to labials, dentals and alveolars. [−anterior] refers to post-alveolars, palato-alveolars, retroflex, palatals, velars, uvulars, pharyngeals.

labial / non-labial : Labial sounds are those where there is rounding or constriction at the lips, one or both lips are the active articulator. [+labial] refers to labial and labialised consonants and to rounded vowels

distributed / non-distributed : "Distributed sounds are produced with a constriction that extends to a considerable distance along the midsaggital axis of the oral tract;

nondistributed sounds are produced with a constriction that extends for only a short distance in this direction" (Chomsky and Halle, 1968). Apical from laminal and retroflex from non-retroflex consonants are then distinguished by this feature.

high / non-high : "High sounds are produced by raising the body of the tongue toward the palate; nonhigh sounds are produced without such a gesture" (Chomsky and Halle, 1968). [+high] refers to palatals, velars, palatalised consonants, velarised consonants, high vowels, semi-vowels.

mid / non-mid : some authors add this feature to deal with vowel systems with four contrastive levels of height. Mid sounds are produced with tongue height approximately half way between high and low sounds.

low / non-low : "Low sounds are produced by drawing the body of the tongue down away from the roof of the mouth; nonlow sounds are produced without such a gesture" (Chomsky and Halle, 1968). [+low] refers to low vowels, pharyngeal consonants, pharyngealised consonants.

back / non-back : "Back sounds are produced with the tongue body relatively retracted; nonback or front sounds are produced with the tongue body relatively advanced" (Chomsky and Halle, 1968). [+back] refers to velars, uvulars, pharyngeals, velarised consonants, pharyngealised consonants, central vowels, central semi-vowels, back vowels, back semi-vowels.

front / non-front : this feature is added to distinct the central vowel, which is [−back] and [−front].

continuant / stop : "Continuants are formed with a vocal tract configuration allowing the airstream to flow through the midsaggital region of the oral tract: stops are produced with a sustained occlusion in this region" (Chomsky and Halle, 1968). [+continuant] refers to vowels, approximants, fricatives. [−continuant] refers to nasal stops, oral stops. Some authors also make a distinction between [continuant acoustic] and [continuant articulatory]. The English nasals [m,n,ŋ] are then [+continuant acoustic] but [−continuant articulatory].

lateral / central : "Lateral sounds, the most familiar of which is [l], are produced with the tongue placed in such a way as to prevent the airstream from flowing outward through the centre of the mouth, while allowing it to pass over one or both sides of the tongue; central sounds do not invoke such a constriction" (Chomsky and Halle, 1968). [+lateral] refers to lateral approximants, lateral fricatives, lateral clicks.

nasal / oral : "Nasal sounds are produced by lowering the velum and allowing the air to pass outward through the nose; oral sounds are produced with the velum raised to prevent the passage of air through the nose" (Chomsky and Halle, 1968). [+nasal] refers to nasal stops, nasalised consonants, nasalised vowels and less-common pre-nasalized stops, nasal glides, nasal fricatives, and nasal trills.

tense / lax : This feature applies to vowel characterization, the traditional definition is that tense vowels present a greater constriction than lax vowels. In some languages, the tense/lax distinction is also made between long/short vowels. A more general definition, states that the tense/lax distinction is related to some kind of strong/weak contrast. In some languages, this distinction is made between more peripheral vowels (closer to the four corners of the vowel quadrilateral) and less peripheral vowels (more centered and/or more mid vowels).

sibilant / non-sibilant : Sibilants are a type of fricative or affricate consonants, resulted from the production of a jet of air through a narrow channel in the vocal tract towards the sharp edge of the teeth. They have then large amounts of acoustic energy at high frequencies, and sound then louder than their non-sibilant counterparts. [+sibilant] refer to the following sounds: [s,ʃ,z,ʒ].

spread glottis / non-spread glottis : "Spread or aspirated sounds are produced with the vocal cords drawn apart producing a nonperiodic (noise) component in the acoustic signal; nonspread or unaspirated sounds are produced without this gesture" (Chomsky and Halle, 1968). This feature is used to indicate the aspiration of a segment. [+spread glottis] refers to aspirated consonants, breathy voiced or murmured consonants, voiceless vowels, voiceless approximants.

constricted glottis / non-constricted glottis : "Constricted or glottalized sounds are produced with the vocal cords drawn together, preventing normal vocal cord vibration; nonconstricted (nonglottalized) sounds are produced without such a gesture" (Chomsky and Halle, 1968). [+constricted glottis] refers to ejectives, implosives, glottalized or laryngealised consonants, glottalized or laryngealised vowels.

voiced / voiceless : "Voiced sounds are produced with a laryngeal configuration permitting periodic vibration of the vocal cords; voiceless sounds lack such periodic vibration" (Chomsky and Halle, 1968). When the segment is characterized as [+voiced], the vibration of the vocal folds occurs concomitantly with its articulation.

Although those features are used to characterize speech sounds, in some situations it makes no sense to use some attributes to some sounds. To characterize vowels as

[+strident] or [−strident], [+lateral] or [−lateral] is something that does not concern the nature of vowels. In the same way, the height of a vowel may be characterized by the features high and low. High vowels are [+high] and [−low], low vowels are [+low] and [−high]. No vowel can be simultaneously [+high] and [+low]; but mid vowels are said to be [−high] and [−low] at the same time.

Some feature combinations are considered impossible. The combination [−sonorant, −consonantal], for example, is said as a physical impossibility since [−sonorant] segments would require a major obstruction in the vocal tract, on the other hand, [−consonantal] says that the obstruction cannot be in the oral cavity. The only conclusion would be to require a constriction of the nasal passages, but nostrils are not sufficiently constrictable, leading to a physical incongruence (Odden, 2005). Another matter that is discussed as an impossibility is whether there are syllabic obstruents segments, i.e. [s̩], [k̩]. It has been claimed to exist in certain dialects of Berber[4] (Odden, 2005), but it is still a controversial matter.

The most important contribution to phonology from the distinctive feature theory is that a set of segments may be analyzed into some features and it is possible to identify classes of segments in rules, creating the notion of *natural class*, as a set of sounds that has certain phonetic features in common and is affected in the same way in the same environment. Natural classes can be defined by the combination of some features. For example, [+consonantal, −syllabic], referring to a set of segments which are simultaneously [+consonantal] and [−syllabic]. In order to create a natural class, made of two or more segments, it is necessary that the number of features used to specify this class to be less than the number of features to specify each element inside this class. The three major class features may be used to define together five maximally differentiated classes, as shown in the table bellow:

	a, i, u	r̩, l̩, m̩	y, w, h, ʔ	r, l, m	s, z, p, b
syllabic	+	+	−	−	−
sonorant	+	+	+	+	−
consonantal	−	+	−	+	+

Further classes are definable by omitting specifications of one or more of these features: for example, the class [−syllabic, +sonorant] includes {y, w, h, ʔ, r, l, m}. As defined by Flemming (2005), "natural classes derive from the nature of the set of markedness constraints. For example, sounds can pattern together as a natural class if they violate markedness constraints in the same environment, so given constraints *XA and *XB, A and B can form a natural class". To the set of speech sounds that form a natural class,

[4]The Berber group of languages is formed by indigenous languages of North Africa west of the Nile. They are spoken mainly in Morocco and Algeria.

it may be given a simpler phonetic categorization. Speech sounds that are in a natural class, usually undergo together in a phonological process. "Sounds do not constitute a natural class just because they share feature specifications, the class must contain all the sounds that have those feature specifications"(Flemming, 2005).

According to the theory of phonological features, phonemes are characterized according to a finite set of features. Those features are called distinctive features for there are no two phonemes that share the same features. If they do so, they are the same phoneme. The features are grouped into categories: major class features, laryngeal features, manner features and place features. Features are usually specified in a binary manner, by assigning a positive value [+] to denote the presence of a feature and a negative value [-] to indicate its absence.

The distinctive feature table proposed for the Brazilian Portuguese is represented in the Figure 8.2. In this representation we assign the color white to the features that we would assign '+' and black to the features we would assign '−'. With this form of representation it is easier to realize which features are shared by some speech sounds.

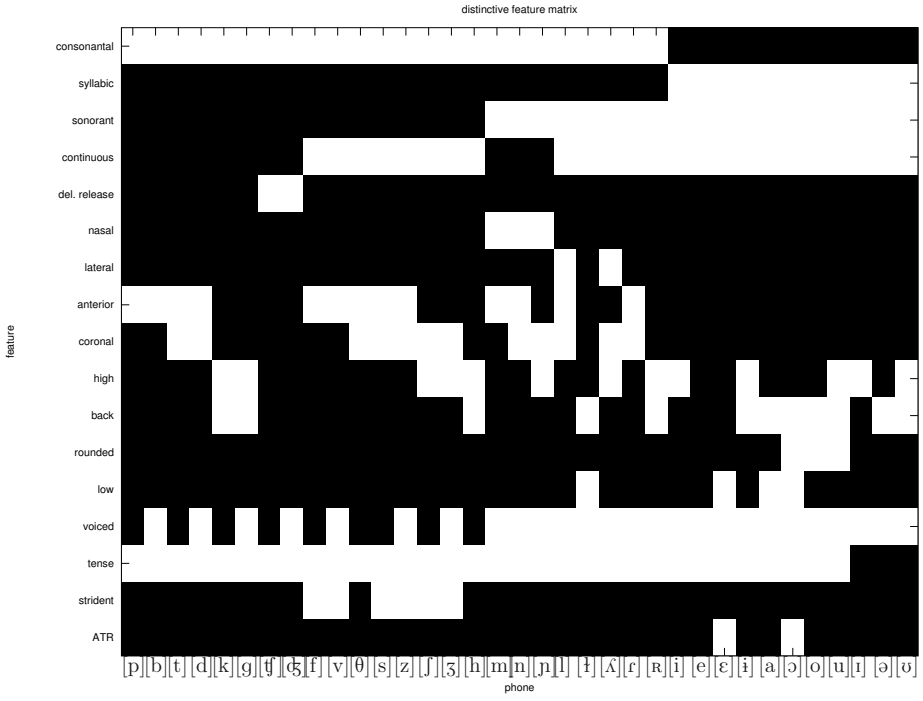


Figure 8.2: Distinctive feature table for the Brazilian Portuguese.

Using this representation of speech sounds into an array of sound features, we may define a distance measure of two segments as the number of features not shared between them (that dissimilarity definition resembles the natural class definition of Flemming (2005)). Using this definition of distance we conclude that the distance between the

segments [k] and [h] is only one, since the only feature they do not share is [continuous], [k] is described as [−continuous] and [h] as [+continuous]. This proposed distance may then be used to create a dissimilarity matrix among all speech sounds in the language repertory. The result of this procedure is shown in Figure 8.3, where the result is normalized so that the maximum distance found is one and the minimum distance is zero. Analyzing this image, it is easy to recognize two big groups: the vowels and the consonants. It is also easy to recognize visually some other groups, as the group of plosives, the group of fricatives, and the group of nasals.

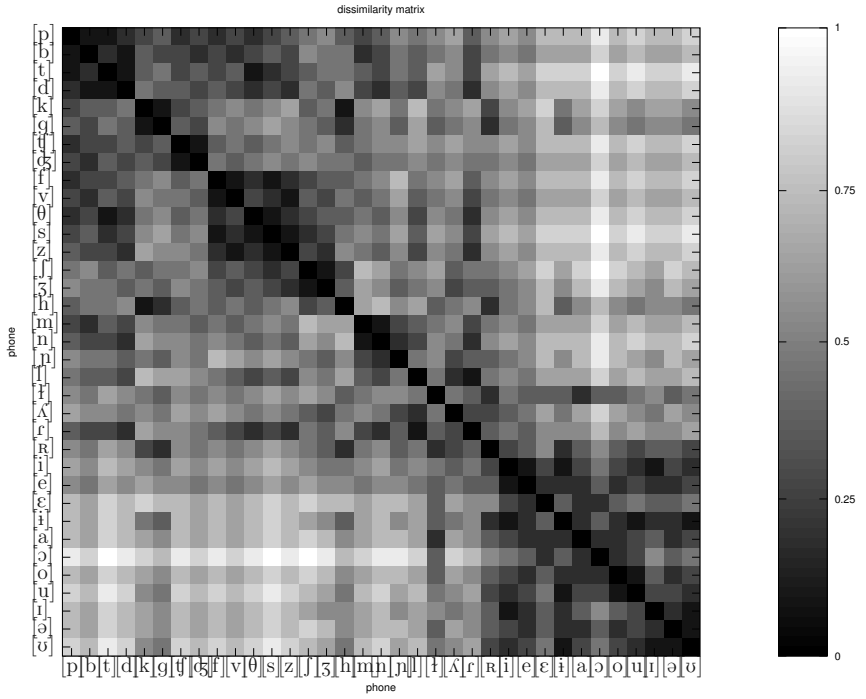


Figure 8.3: Dissimilarity matrix for the speech sounds in Brazilian Portuguese.

With such a matrix of dissimilarities at hand, we might perform the multidimensional scaling (MDS) of the data. The idea of MDS was first introduced by Young and Householder (1938) and it consists of finding a representation of objects in a vector space such that the distance between those representations is in accordance with the dissimilarities presented by the input matrix. "Multidimensional scaling, then, refers to a class of techniques. These techniques use proximities among any kind of objects as input. A proximity is a number which indicates how similar or how different two objects are, or are perceived to be, or any measure of this kind. The chief output is a spatial representation, consisting of a geometric configuration of points, as on a map. Each point in the configuration corresponds to one of the objects. This configuration reflects the 'hidden structure' in the data, and often makes the data much easier to comprehend" (Kruskal and Wish, 1978).

It is important then to choose a suitable metric which leads to a meaningful description of a data space, for wrong descriptions of facts may lead to false results and wrong interpretations. The distinctive feature theory provides a way of characterizing speech sounds based on articulatory, acoustical and perceptual attributes (Chomsky and Halle, 1968). In this theory, there is a unique representation of each speech sound based on presence or absence of features. This paper uses the theory of distinctive features to create a measure of dissimilarity: a distance measure between two segments defined as the number of features that they do not share.

The input for an MDS method is a dissimilarity (or similarity) matrix Δ:

$$\mathbf{\Delta} = \begin{pmatrix} \delta_{1,1} & \delta_{1,2} & \dots & \delta_{1,N} \\ \delta_{2,1} & \delta_{2,2} & \dots & \delta_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{N,1} & \delta_{N,2} & \dots & \delta_{N,N} \end{pmatrix} . \tag{8.1}$$

The dissimilarity (distance) or data value connecting object i with object j is represented by $\delta_{i,j}$.[5]

The output of an MDS method is a set of N R-dimensional vectors representing the objects (or stimulus) subjected to the current study:

$$\begin{aligned} \mathbf{x_1} &= (x_{1,1}, \dots, x_{1,R})^T \\ \mathbf{x_2} &= (x_{2,1}, \dots, x_{2,R})^T \\ &\vdots \\ \mathbf{x_N} &= (x_{N,1}, \dots, x_{N,R})^T \end{aligned} \tag{8.2}$$

To calculate the MDS, we should start by calculating the dissimilarity matrix $\mathbf{D}$ (where, in this case, $\mathbf{D} \approx \mathbf{\Delta}$) of distance between samples of our database. We build then a matrix $\mathbf{A}$ such that:

$$[\mathbf{A}]_{i,j} = a_{i,j} = -\frac{1}{2} d_{i,j}^2 . \tag{8.3}$$

A matrix

$$\mathbf{B} = \mathbf{HAH} \tag{8.4}$$

[5]In many situations there may be no effective difference in meaning between $\delta_{i,j}$ and $\delta_{j,i}$, and there may be no meaning at all for $\delta_{i,i}$, so that the data values may not form an entire matrix, but only part of one.

is calculated, where $\mathbf{H}$ is given by

$$\mathbf{H} = \mathbf{I} - N^{-1}\mathbf{1}\mathbf{1}^T \ , \tag{8.5}$$

which is positive-semidefinite by construction. The matrix $\mathbf{I}$ used is the $N \times N$ identity matrix, and $\mathbf{1}$ is an N dimensional all-ones vector.

The matrix $\mathbf{B}$ is usually a positive-semidefinite matrix (if not, we may add a small constant and make it positive-semidefinite), so that singular value decomposition (SVD) applies:

$$\mathbf{B} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T \ . \tag{8.6}$$

$\mathbf{B}$ is a matrix of rank p, then $N - p$ eigenvalues of $\mathbf{B}$ are null and we have

$$\mathbf{B} = \mathbf{V_1}\mathbf{\Lambda_1}\mathbf{V_1}^T \ , \tag{8.7}$$

where $\mathbf{\Lambda_1} = diag(\lambda_1, \ldots, \lambda_p)$ and $\mathbf{V_1} = [v_1, \ldots, v_p]$. As $\mathbf{B} = \mathbf{X}\mathbf{X}^T$, $\mathbf{X}$ is given by

$$\mathbf{X} = \mathbf{V_1}\sqrt{\mathbf{\Lambda_1}} \ , \tag{8.8}$$

and we have then calculated the matrix $\mathbf{X}$, which is the representation of the original data in a new vector space where each dimension is a principal component.

Using the dissimilarity matrix created and explained above, an MDS procedure is performed and a representation in two dimensions is shown in figures 8.4 and 8.5. The first plot, represents the vowels of Brazilian Portuguese in a two dimensional space. It is important to note that this 2D plot represents 65% of the data variance, meaning that more dimensions would be required to take into account all the information of the data. Anyway, analyzing the result, we can easily recognize the dimensions used in the usual vowel diagram (shown in Figure 8.7) where the main distinction is made in terms of *high* and *low*, and *back* and *front*). The dimensions of the MDS representation are in fact a combination of the features *high-low*, and *back-front*. The same sort of analysis may be performed on the MDS representation for the consonants. It is shown in Figure 8.6.

Considering one more dimension, we are capable of creating a representation in three dimensions where now we have 61% of the data variance explained by this plot. With this new plot it is possible to see a better separation of some speech sounds, the fricative and the plosive sounds are now represented further apart compared to the first representation in two dimensions.

Figures 8.9, 8.10, 8.11, 8.12, are analogous to figures 8.3, 8.4, 8.5, 8.8, but now the distance metric used is another one. In this new metric we add null if two speech sounds share a feature with a '+' attribute; if both share a '−' we add 1/3 to the distance

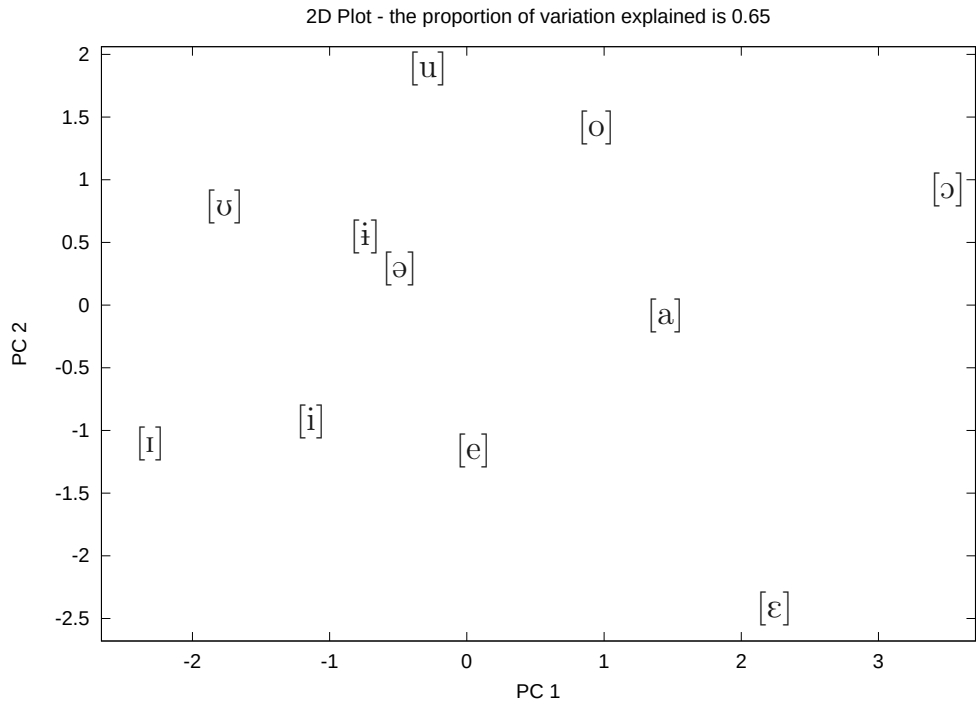


Figure 8.4: 2D MDS result for the vowels of Brazilian Portuguese.

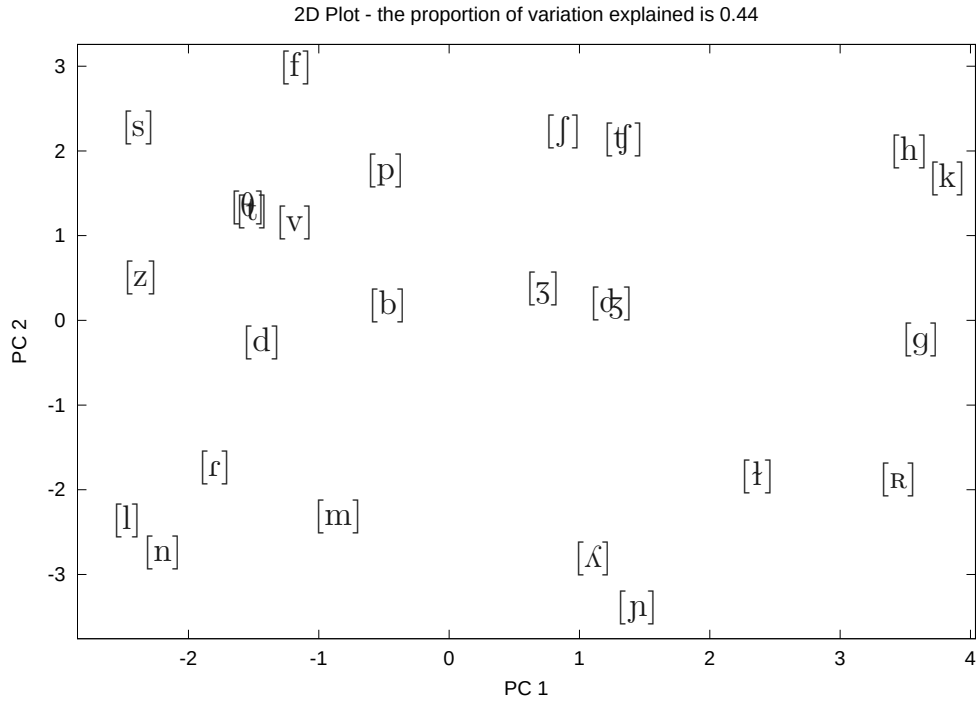


Figure 8.5: 2D MDS result for the consonants of Brazilian Portuguese.

measure, because a '−' attribute may be used by something to what this attribute does not make sense; and we add 2/3 to the other possible situations (['+','−'] and ['−','+']), for the same reason. The results show that there is only a slight change in the MDS

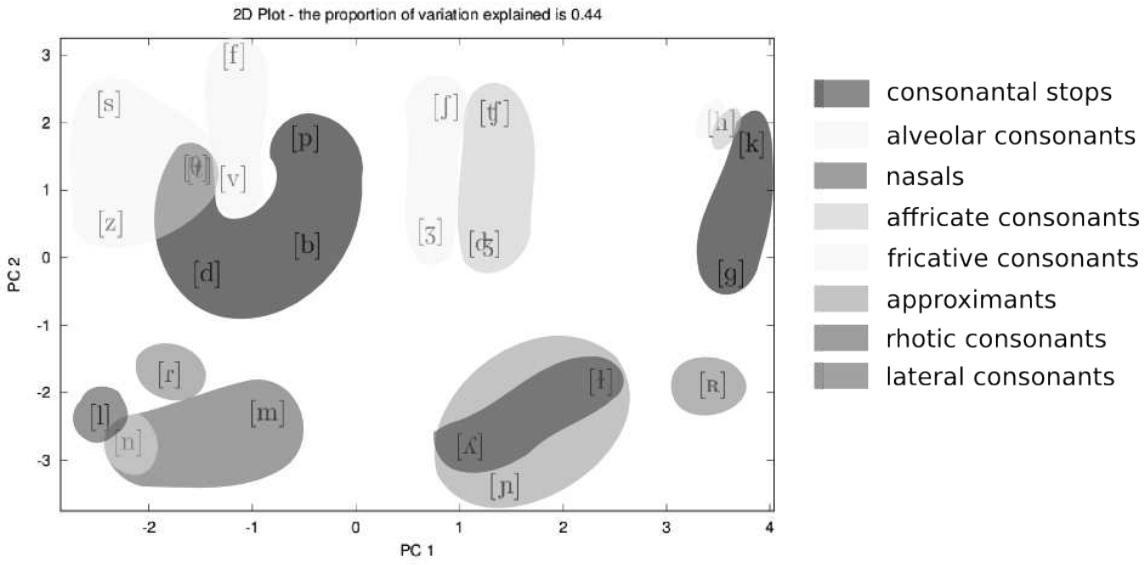


Figure 8.6: 2D MDS result for the consonants of Brazilian Portuguese.

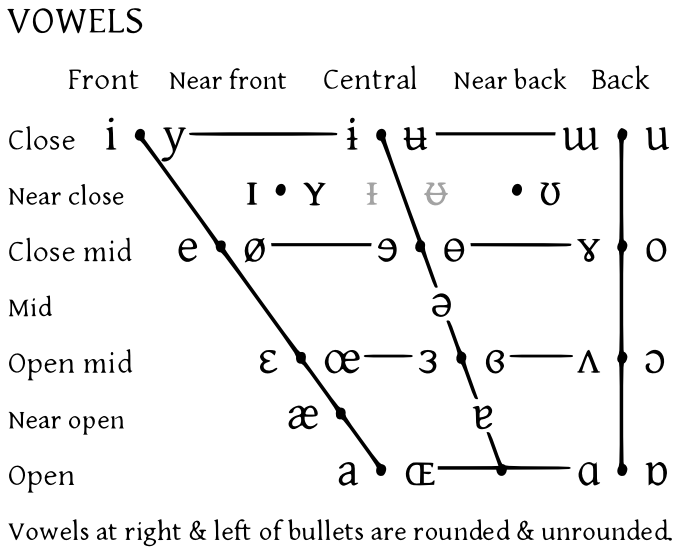


Figure 8.7: Vowel Diagram from the IPA table.

representation found.

The distinctive features are used not only to characterize and define phonemes, but also to formalize class of segments and phonological rules. Every specification of a feature defines a class and the generality of a class is inversely proportional to the number of features used to define it. If we consider only one feature [+syllabic], it refers to a class of all syllabic segments, that means all vowels {ɪ, i, ɛ, e, a, ɔ, o, ʊ, u, ə, æ, ĩ, ĩ̞, ɛ̃, ẽ, ã, ɔ̃, õ, ʊ̃, ũ, ə̃, æ̃}. As we add features to describe a class, we narrow down the possibilities and get a less general class. Considering now the class of segments described by [+syllabic, −nasal], the class will be now restricted to {ɪ, i, ɛ, e, a, ɔ, o, ʊ, u, ə, æ}. Adding one

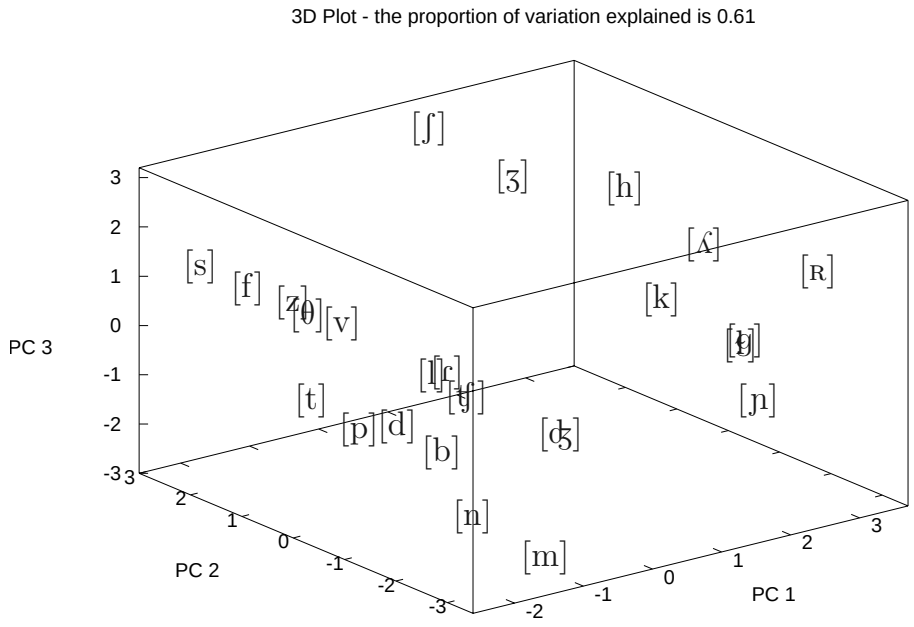


Figure 8.8: 3D MDS result for the consonants of Brazilian Portuguese

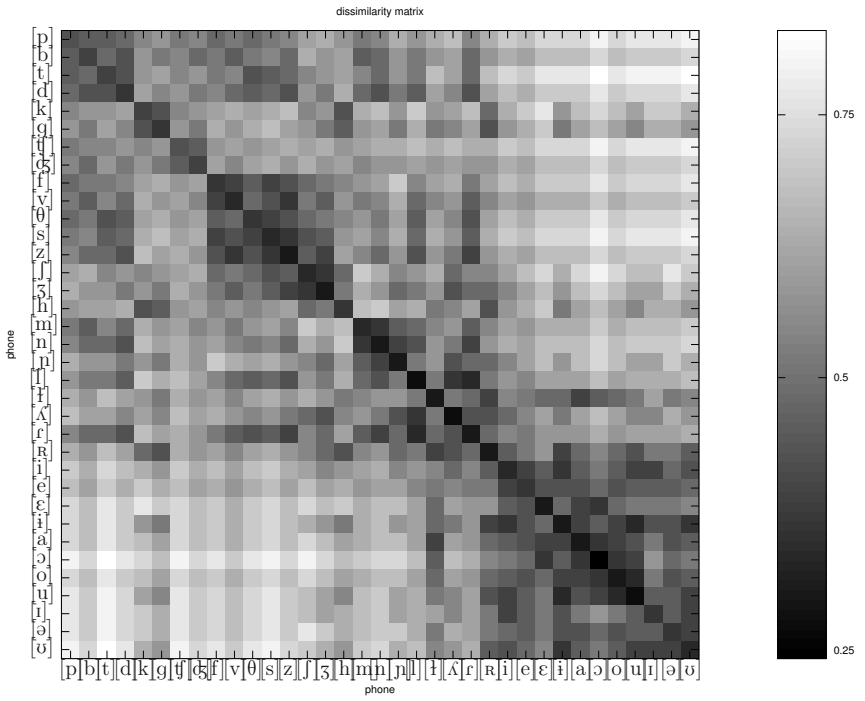


Figure 8.9: Dissimilarity matrix for the speech sounds in Brazilian Portuguese (using the second proposed metric)

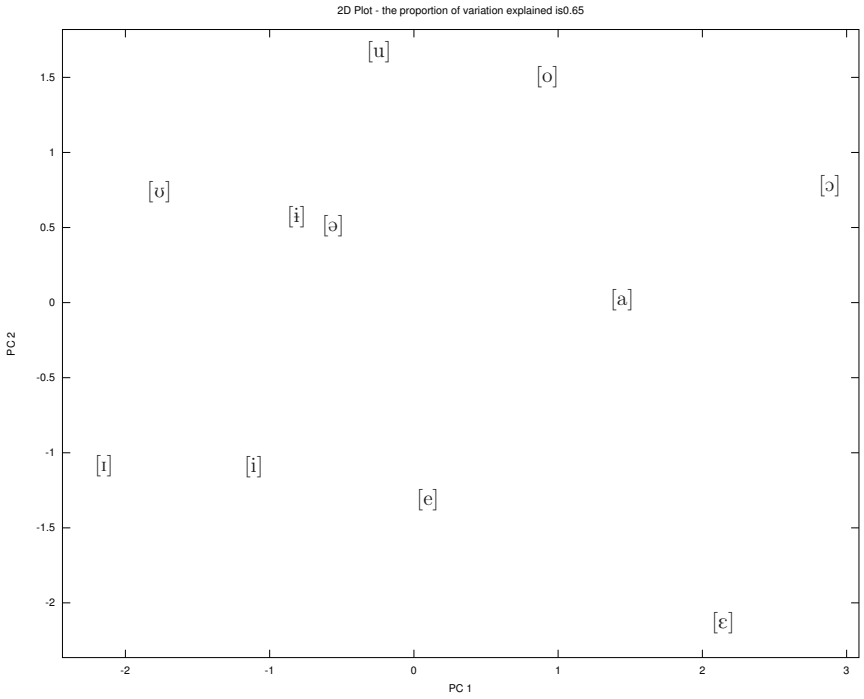


Figure 8.10: 2D MDS result for the vowels of Brazilian Portuguese (using the second proposed metric).

more restriction to our class definition [+syllabic, −nasal, +round], we get the following {ɔ, o, ʊ, u}. We can narrow our class description by adding one more feature [+syllabic, −nasal, +round, +high], and now we have {ʊ, u}. Finally, adding one last description [+syllabic, −nasal, +round, +high, −tense] we can restrict our class to a one-element-class {ʊ}. The principle used in phonology to describe phonological rules is to choose simpler descriptions (which use fewer features) to describe a phonological rule.

The feature theory provides an upper limit to the number of possible phonemes, that is 2^n, where n is the number of features used to describe the sounds of a language. For the previous example of Portuguese, we used $n = 17$ features to describe all speech sounds. That makes a total of $131,072$ possible combinations of features, which is our theoretical upper limit. This number is a lot greater than the number of phonemes in our structuralist analysis. We must note that some feature combinations are physically impossible to be realized, so the number of speech segments is in fact much smaller than $131,072$. For example, all combinations of [+high,+low] must be excluded, since it makes no sense, as the tongue cannot be contradictorily raised and lowered at the same time. Only this analysis excludes $2^{15} = 32,768$ possible combinations from the total. Another case of impossible articulation is the combination of features [+consonantal, −high, −back, −anterior]. The feature [−back] specifies a segment with a place of articulation in front of the velar po-

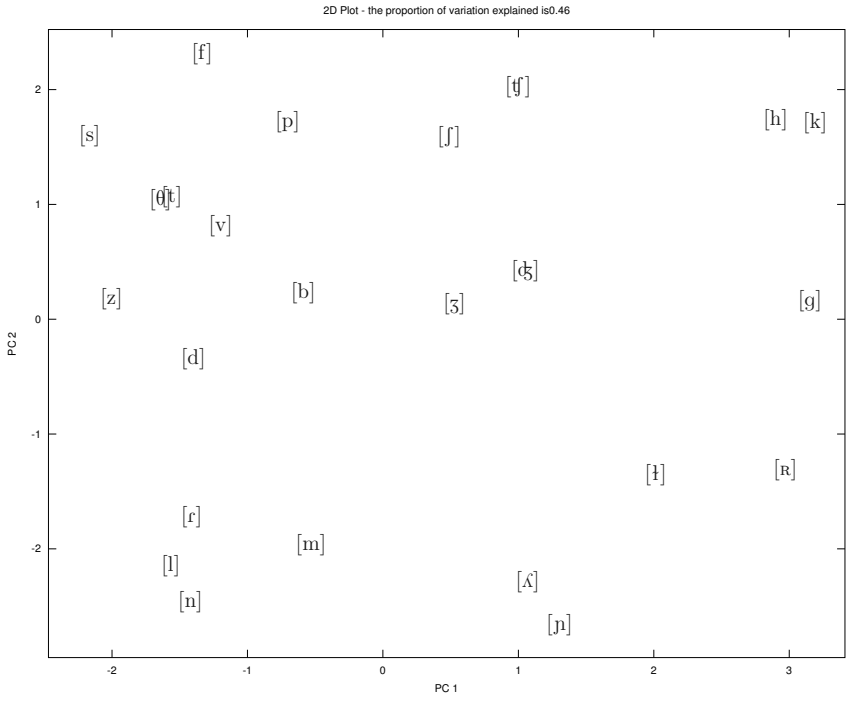


Figure 8.11: 2D MDS result for the consonants of Brazilian Portuguese (using the second proposed metric).

sition. The feature [−anterior] requires a place of articulation behind the alveolar ridge. The feature [−high] implies that it cannot be a palatal. The only possible segment for the combination of features [−high, −back, −anterior] leads us to [e] as the only possible segment with the specified features. If we require another feature [+consonantal], then we conclude it cannot be [e], since it is a vowel and then [−consonantal]. So, all combinations with [+consonantal, −high, −back, −anterior] are impossible speech sound realizations.

Although the features description may be accurate to describe a great deal of speech sounds, it has some drawbacks. Retroflex consonants, for example, are described by [−anterior, +coronal, −distributed]. As pointed out by Odden (2005), a contrast cannot be expressed between apical and sublaminal retroflex, as found in Hindi versus Telugu, according to this theory of features. "Similarly, the differences attested in the phonetics of [u] and [ʊ] across languages are never found within a language. In a single language, the maximal contrast is between two such vowels, governed by the feature tense (or ATR). The fact that such differences exist at the phonetic level between languages, but are never exploited within a single language as a way to distinguish words, is an example of the difference between phonetic and phonological properties" (Odden, 2005).

The original set of features originally proposed by Chomsky and Halle (1968) may not be sufficient to make good description of every language. Suppose there is a language

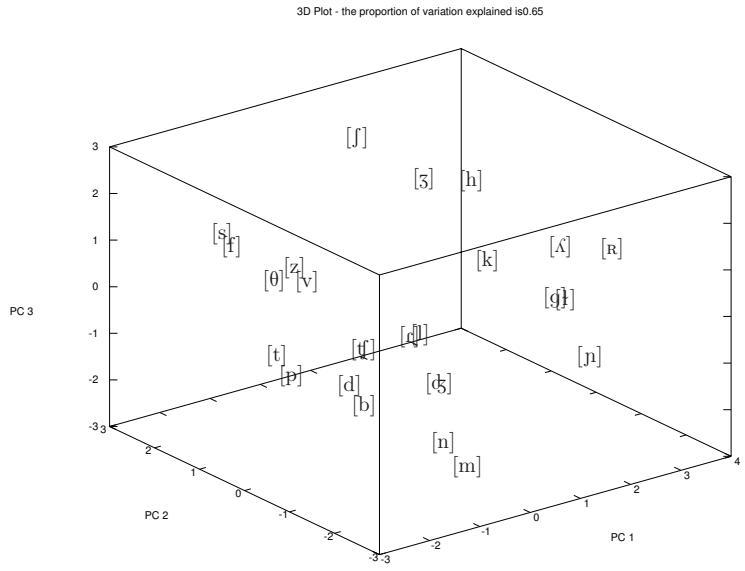
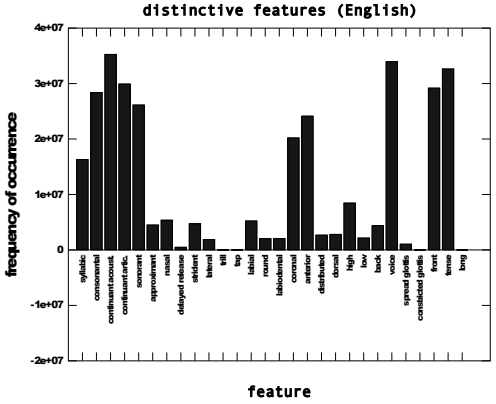


Figure 8.12: 3D MDS result for the consonants of Brazilian Portuguese (using the second proposed metric).

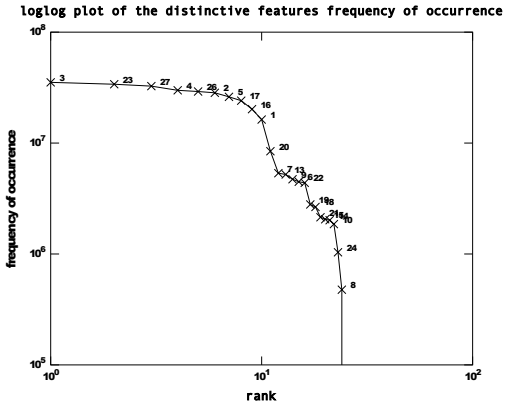
which makes contrast between regular and sublingual retroflex consonants. A [sublingual] feature should be added to account for distinctions of such kind. The addition of a new feature should be made only when there is compelling evidence to do so and when there is no other way to explain the speech sound found in this language. A classical case is the feature [labial], which was not taken in account by the theory of Chomsky and Halle (1968), and was added to make phonological rules better formalized (Odden, 2005).

Considering the distinctive features presented in this chapter and the frequency of occurrence of phones analyzed in the last chapter, it is possible to join both pieces of information and introduce a new approach. In the last chapter, the English phones were ranked according to their frequency of occurrence. We may use that frequency information and the description of each phone into a set of distinctive features, and then we may infer the frequency of occurrence of distinctive features. That is presented in Figure 8.13a, and the log-log plot of these features sorted by their frequencies is shown in Figure 8.13b.

We might suppose that two phones occur more frequently in a diphone because they are more contrastive, and then a dissimilarity measure between two diphones may be built using the co-occurring frequency. Considering also the dissimilarity based on distinctive features (the number of features two phones do not share), we may compute one dissimilarity as the product of this number and the frequency of co-occurrence. The result is



(a) Frequency of occurrence of distinctive features in English.



(b) Distinctive features in English ranked by their frequencies of occurrence.

Figure 8.13: Analysis of the Distinctive Features in English.

presented as a matrix in Figure 8.14, where black stands for no dissimilarity and white for maximal dissimilarity. Using that result into an MDS analysis, we get the 2D representation displayed in Figure 8.15. Unfortunately we might not get much information from it, since most of the phones are collapsed into a small region of the PCs space.

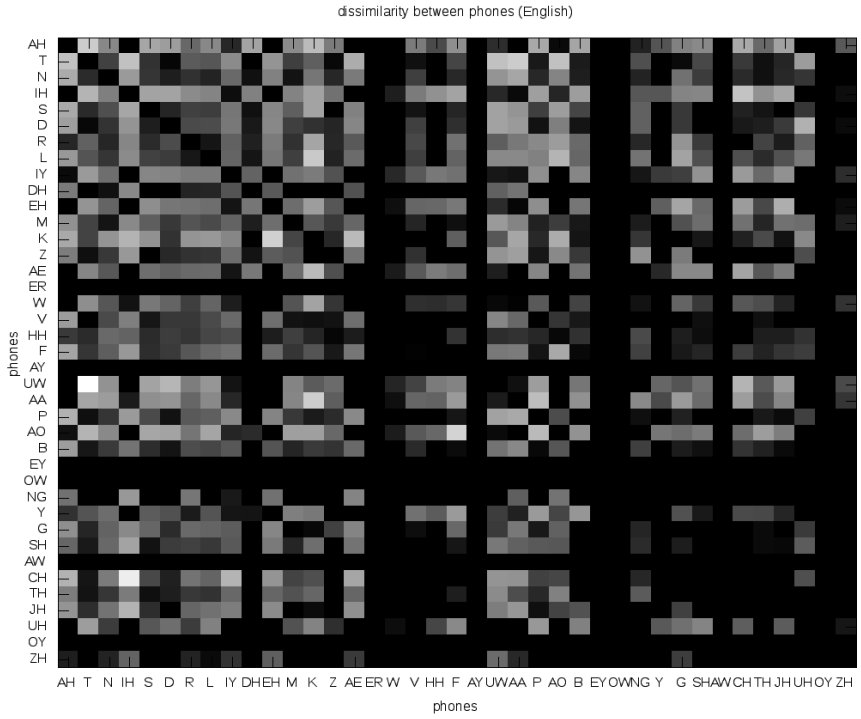


Figure 8.14: Dissimilarity between phones (the frequency of co-occurrence is used as a dissimilarity measure).

Using the frequency of occurrence of these features, we may build a dissimilarity mea-

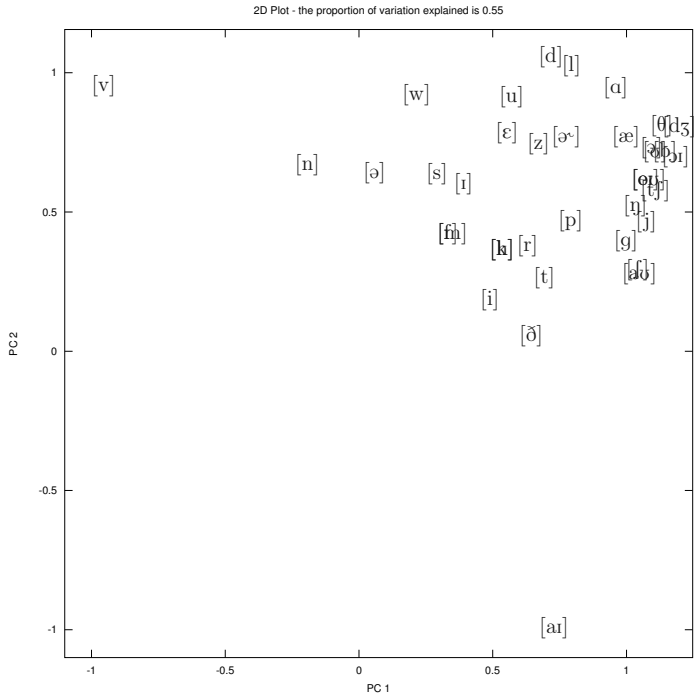


Figure 8.15: MDS result for the English phones, considering the co-occurrence frequency as a dissimilarity measure.

sure based on the frequency of co-occurrence of features in a diphone. The dissimilarity of two features is measured on the frequency of their co-occurrence in diphones. This definition leads to the dissimilarity matrix between features, which is presented in figure 8.16. This dissimilarity measure may be used as the input to an MDS. The result is shown in figure 8.17. An analysis of this output is quite interesting. We observe that the features [syllabic] and [consonantal] are placed at the greatest distance possible, they are almost exclusive features (the only exceptions are glides). Observing that 12 features are distributed around the PCs space and the 16 remaining are pile up on a small peace of the PCs space, we might suppose that it is due to the distinctive capability of the features in the language under observation.

The analysis that was carried out is based on distinctive features which were selected from the point of view of speech production, but do not necessarily reflect perceptual aspects of the human capability to distinguish speech sounds. In order to take it into account, the same procedure can be made using a subjective dissimilarity measure. A group of subjects is asked to listen to some speech stimuli grouped into triads. For each triad, the subject must elect the most similar and the most dissimilar pair. Based on the results of all subjects, a dissimilarity matrix can be created, and an MDS procedure is used to find a spatial representation of the stimuli. The presented stimuli may be free

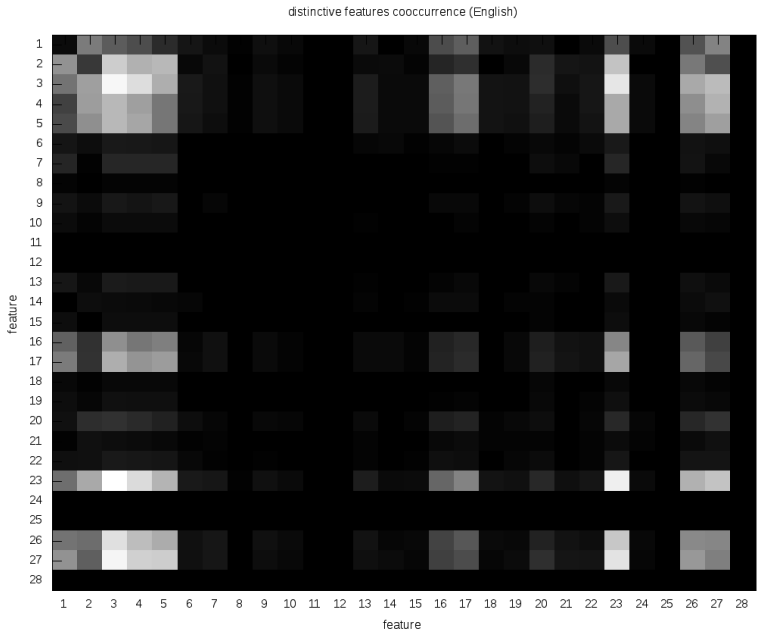


Figure 8.16: Frequency of co-occurrence of distinctive features in diphones.

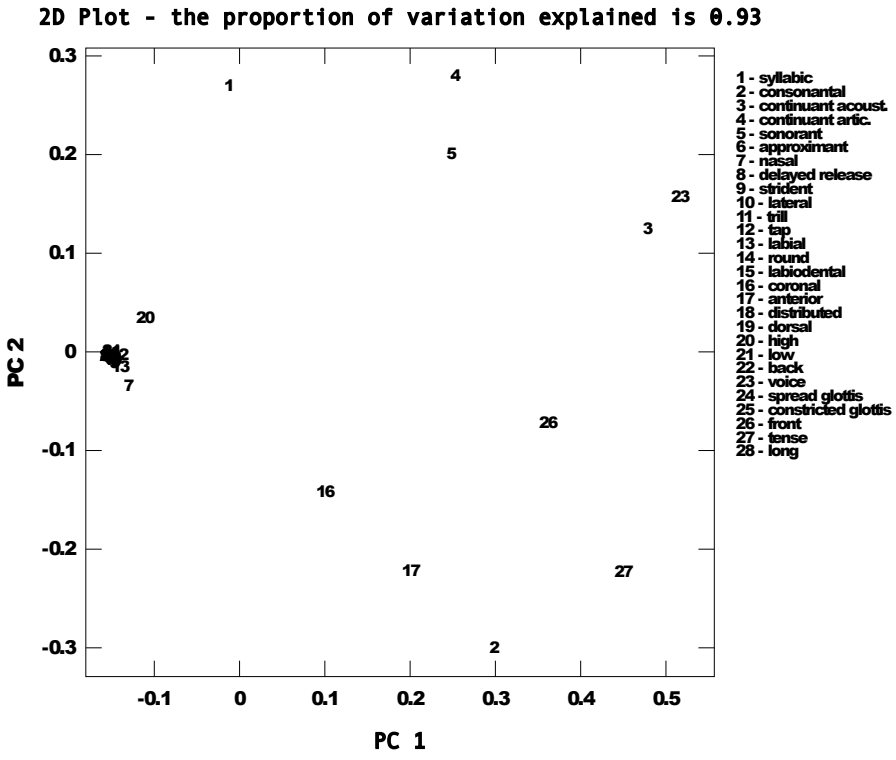
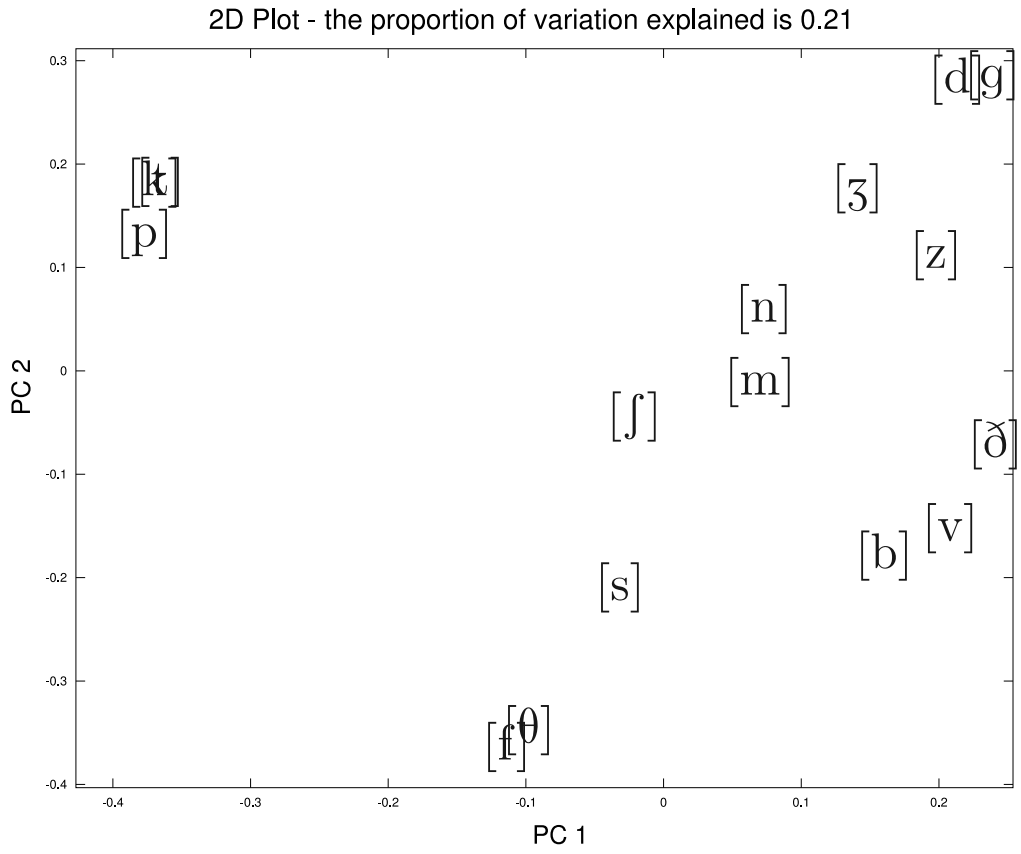


Figure 8.17: MDS result for the distinctive features in English.

vowels in order to create a dissimilarity measure only for vowels, or CV syllables, where we can choose different consonants and use them to create a consonantal dissimilarity

matrix.

Miller and Nicely (1955) presented an experiment where a series of stimuli, sixteen English consonants, were presented to subjects who were forced to guess them at every sound. The stimuli were presented under different filtering (low and high pass) and various signal-to-noise ratios (SNR). The results were given as a confusion matrix. The results from Miller and Nicely were used to create a dissimilarity matrix and subsequently a MDS was performed, in order to compare these results with those achieved through the feature theory approach. Figures 8.18a and 8.18b present MDS results from the distinctive feature data for the English consonants. In the Figures 8.18a, it was used a combination of 17 different dissimilarity matrices that were given by Miller and Nicely (1955). We may see some grouping, for example, the voiceless stops on the upper left. The more interesting is to note that voicing and aspiration are presented as the major components on the placement of speech sounds this Figure. In the next Figure is presented the plot result when only the filtered data between 200 and 300 Hz and with 12dB SNR is used. We see the formation of tree groups: nasals, voiced and voiceless consonants. We clearly realize that voicing is a quite good to differentiate speech sounds and that information is well represented in this tight band.



(a) 16 English Consonants - variation explained is 0.21. Results from a combination of 17 dissimilarities matrices.

Figure 8.18: MDS results using the data from Miller and Nicely (1955)

9

Conclusions

In this study we have focused on applying a statistical analyses on language use data, in special investigating some already know quantitative linguistic laws. Initially we have performed a horizontal analysis across different languages, using the UPSID. Afterwards we have attained our attention to the patterns of use of only one language: English, performing the in multiple linguistic levels. We have also the classical Information Theory to carry out a systematic inquiry to examine language under this perspective, in order to understand the trade-off between efficiency in information transmission and complexity of the system. These analysis have shown important to achieve a quantitative comprehension of linguistic concepts which are already well known and described in the literature. Such study is important to understand the processes underlying language use and evolution, what is necessary to create a better model that might be applied to create technological artifacts that truly exhibit human-level communication capabilities.

The understanding of language origin and evolution requires a deep study and identification of what are the universal features that drive language evolution. We adopt the usage-based theory of language in which its cognitive organization is a product of the speaker experience, rather than a set of abstract rules and structures that indirectly relates to the language experience. Grammar is based on usage, contains many details of probabilities of occurrence and co-occurrence, and creates a dynamic network from categorized instances of language use. Frequency of use and its consequent cognitive or-

ganization have impact on language structure and usage. Speed of access related to token frequency, priming, the process of grammaticalization and the morphological and phonic characteristics of high and low frequency words are examples of the impact frequency has on language and cognition (Bybee, 2001, 2003, 2006; Ellis, 2002). Studies (Saffran et al., 1996a, 1999; Saffran and Wilson, 2003) show that human track patterns and statistical regularities in artificial grammars even when there is no correspondence to meaning or communicative intentions.

The idea of 'universals of grammar' may be seen as a by-product of the emergent phenomena of grammar, what is dependent on idiosyncratic localized interactions and the process of spreading and shifting through populations and time. From the basic principle that usage is a driven factor to shape language, we expect to observe common patterns and regularities in different languages. The UPSID database is very important to analyze the type of segments used by different languages in the world to build their speech segments repertoire. From the observation of the data in this database we could realize that some speech segments are very common among languages, while others are quite rare. The phone [m], for example, happens in 94.2% of the languages in that database, while 427 speech sounds appear just once among all 425 languages.

It is a common place to use more consonants than vowels (90% of the languages use at most 6 times more consonants than vowels). The median number of vowels used by languages is 4, and the median number of consonants is about 18. We could observe that just a few phones have many co-occurring segments in their languages and are also found in different languages. It could give us a hint that they work like wildcards. Although some results might suggest that there is a binding force between voice and unvoiced counterparts, specifically, the existence of an unvoiced consonant would make it more probable the existence of its voiced counterpart (not the other way around), it seems this is false, since the probability of a voiced segment given its voiceless counterpart is 0.406, and the probability of a voiceless segment given its voiced counterpart is 0.415. Moreover, the frequency of occurrence of voiced and unvoiced segments in UPSID is not so different. Voiced segments appear on 3.8% of the languages and unvoiced on 3.1%. Another feature is that languages that use fewer segments to build their speech repertoire tend to use mostly common segments, and languages that use many speech segments, tend to use rare phones to build their repertoire.

Our first analysis consisted on observing the segments used by various languages around the world to build their speech repertoire. As we understand language as a dynamic self-organized system, we believe it is important to investigate how those segments are used, how they build larger structures, what are the statistical characteristics of languages. We believe that there might resemblances on the way languages use their

segments and in the way they structure themselves through usage. This type of analysis would require a great amount of work, and therefore we have restricted our analysis to the linguistic patterns observed and their mathematical models. Those patterns observed across different languages or phenomena are also called *laws*.

The ubiquity of scaling laws in nature suggests that complex systems organize themselves, they are not just a collection of random independent realizations, but exhibit similar patterns and relations that are observed in different scales. "Scaling laws (...) pervade neural, behavioral and linguistic activities"(Kello et al., 2010). The ubiquity of normal distribution in nature is also well known. Although in the past it was not satisfactorily understood, it became clear when the central limit theorem showed that random independent variables always combine to produce, in the limit, a normal distribution. Complex systems are more than mere collections of independent random variables, they have interdependent effects and behaviours, and therefore might not be described by random distributions.

Scaling laws are also present in many cognitive activities. Steven's power law (Stevens, 1957) proposes a relation between the magnitude of a physical stimulus and its perceived intensity, which are related by a power law in the form $\Psi(I) = kI^{\alpha}$, where I is the magnitude of the physical stimulus, α is the relating exponent and k is the proportionality constant (which depends on the type of the stimulation). Steven's power law is also known as a generalization of Weber-Fechner law to a wider range of sensations. This law shows the existence of similar relations among variables at different scales, which means that there is a scale invariance in perception. These observed scaling laws are hypothesized to reflect the maximization of sensitivity and dynamic range of sensory systems (Copelli and Campos, 2007) In the case of motor control systems, it is argued to be used to minimize the effects of errors due to noise in the motor system (Harris and Wolpert, 1988). The idea of a time dynamic scaling memory retrieval, proposed by Brown and Neath (2007), is also consistent with the ubiquitous observations of scaling laws in cognition.

Some researchers claim that Zipf's law is shallow, since it is a result of choosing rank as the independent variable and for that reason even random generated texts present a Zipf pattern. However we have presented some arguments claiming that the intermittent silence model is shallow itself and it does not hold all features present is a natural communication process. A visual analysis of the log-log plots of frequency of occurrence versus rank creates a first evidence. The Zipf model presupposes a monotonically decreasing sequence of values for the frequency of occurrences, however for a random generated text, all words with the same length share the same probability of occurrence, creating a stair case pattern. This is not observed in natural texts, expect for the poorly sampled words, which appear at the end of the tail of the decreasing curve. Another aspect that

makes natural and random texts distinct is evidenced by the histogram of words length. Random generated text present a geometrically distributed frequency of occurrence for words length. The same is also present when considering the number of words for a given length. The number of words, as well as the frequency of occurrence of words, suffer a steeper decrease when approaching very short word lengths. A maximum is found for words which length is equal to 7 letters, or 5 phones, or 1 syllable, when regarding the number of existing words; or a maximum frequency of occurrence is found for letters of length equals to 3 letters, or 2 phones, or 1 syllable (depending on your unit adopted to measure a words length). Random texts present an exponentially decreasing number of words as the words length increase. The same is observed when regarding the frequency of occurrence. The plots of natural texts also show a much steeper decrease in the number of words and its frequency as the length increases, in comparison with random texts.

We have simulated the creation of random texts, generating random symbols. We used three different approaches: uniform distributed symbols; symbols with the same probabilities as what is observed in natural texts; symbols generated by a Markov model with the same transitions probabilities from the natural text. Only the random text created by a Markov model gets close to the observed pattern of 2-grams and 3-grams in real texts. The Markov model was also able to produce words with the same behaviour found in natural texts, a higher word length around 2-3 symbols long words.

The inverse Zipf plot is another clue leading to the fact that random texts do not present true Zipf distribution. The occurrence of *hapax legomenon* on random texts is greater than in natural texts. More over, random texts seems to present a more ragged pattern in the region of high occurrence words, "real texts fill the lexical spectrum much more efficiently and regardless of the word length, suggesting that the meaningfulness of Zipf's law is high"(Ferrer-i-Cancho and Solé, 2002). Other arguments are presented, such as the evaluation of consistency of ranks, indicating that Zipf's law might in fact be a fundamental law in natural languages (Ferrer-i-Cancho and Gavaldà, 2009; Ferrer-i-Cancho and Elvevåg, 2010).

The exponent parameter in the Zipf distribution is important in determining the statistical properties of that distribution. In order to compare data that follow a power law relation, we need to find the best fit to the data. We have presented the usage of the maximum likelihood approach to determine the exponent parameters that gives the best fit. We have derived the equations which lead to a root finding problem, for which the solution is the maximum likelihood estimated parameter. The same procedure is used to find the exponent and flattening parameters of a Zipf-Mandelbrot model. This procedure was tested using real text data and the results show a good performance.

A definition of a distance measure between two different phones was proposed, consist-

ing on the number of distinctive feature not share between these phones. This definition is in accordance with the natural class definition proposed by Flemming (2005). We argue that phones are chosen to built higher order structures according to their resemblances and distinctions. We analyzed the formation of triphones according to this distance measure and we concluded that, more frequently, the triphones are build using phones that have an intermediate distance between them. Too much similarity between consecutive phones must be avoid, in order to create easy distinction between them, and so too much disparity must also be avoided, to ensure easy in the articulation. This reflects once again the trade off between the speaker effort and the listener effort. We have also observed that the first pair of phones in a triphone tends to present more distinctiveness than the second pair of phones.

Menzerath's law states a relation on the usage of tiles to create a whole. According to this law, the length of the whole is inversely proportional to the length of the parts. We considered the duration of words utterance to estimate an average duration in the pronunciation of word's parts (syllables and phones). The results show that as the word length (number of syllables of phones) increases, the average duration of the parts decreases. We also observe this type of behaviour as we analyze the relation between sentences length (number of words) and words length (number of syllables or letters).

As we consider language as a mean to communicate, exchange information, it is important to analyse it under the information theory point of view. We want to characterize a source, which produces symbols according to a Zipf distribution, according to the information content generated. We used the same steps proposed by Grignetti (1964) to derive an estimate of the entropy of such a source. The exponent parameter of the distribution was shown to be a sensitive parameter to characterize the source regarding its information production. The estimates obtained, when compared with real data are shown to be slightly different, what shows that the proposed approximation is good. Small discrepancies are explained by numerical errors, small deviation on the estimated distribution parameters, and the undersampling on the data, which is represented as a staircase pattern on the low frequency region.

Cristelli et al. (2012) proposes that Zipfian distributed data hold coherence among them. We present the distortion created when the data from a certain range of ranks are extracted from the group. This process may create a distortion on the frequency-rank relation. There will be no distortion if the extraction process itself only on high rank types. There will be distortion on the low rank types if low rank types are extracted. The extraction of the very high frequency types is called as the New York's effect. Cristelli et al. (2012) argues that when there is coherence among data, you may gather different data and still the Zipf's law will still hold. But if you gather data from different sources

with distinct characteristics, even if they are coherent by themselves, when they are added, the new global data will not be coherent itself, and the Zipf's law will not hold anymore. A simple example of the concatenation of different random data, created by different probability distributions, show that the same pattern is also observed after the random data is concatenated. This could contradict what was expected by Cristelli et al. (2012), or could corroborate that random data are not truly Zipfian distributed, and the Zipf like pattern was once mode generated as a byproduct of selecting rank as the independent variable.

Leijenhorst et al. (2005) shows that a source which is distributed according to a generalized Zipf's law will present a Heap's law behaviour, there will be a sublinear relation between the text length and the vocabulary size. We present a calculation to show that any source with any distribution will present a monotonically growing lexicon size which converges to the underlying lexicon size. The sublinear lexical growth pattern is exemplified using a few examples from the real world.

In order to analyze the patterns of a language under different scales, analyzing the occurrence of phones and formation of higher order structures, we believe the usage of the Feature Theory might be providential since it give us a direct and formal way to quantify a phone characteristics making it possible to create a distance measure between two phones. We propose to measure the distance between two phones by the number of distinctive features not shared by them. By means of a multidimensional scaling we create a representation of the phones in a vectorial space where the given linear distance will provide the best match with the proposed distance measure. The graphical representation shows that this measure might be a meaningful choice. We believe that its usage concomitantly with frequency of occurrence might give us a better hint on the structure and organization of languages.

10

Further Work

A study on language phonology, where the focus is on the statistics of language, was presented here. Language, as a social tool, is created and driven by society and its changes. The way languages are structured regulate their usages and their usage shapes languages structures. Languages are then self-organizing feedback systems. In order to understand how this complex system words, it is also necessary to investigate quantitatively the patterns observed in language usage.

Some analysis developed here focuses only on English due to the wide spread usage in phonology study, what makes it easy to find tools and references. Here we used a text database to build a list of words ranked according to their frequencies of occurrence. This list was later transcribed through a pronouncing dictionary, which gives us a ranked list of spoken words represented by their phones. It is assumed that the frequency of occurrence of phone, diphone, etc. in a language is similar to the frequencies found by means of this ranked list. The occurrence and co-occurrence of phones, the occurrence of phones in a word, the phonemic length of words, etc. were analyzed. Other questions might still be answered using this approach, and other analysis might also be proposed.

In order to acquire a better insight in this subject, it is necessary a better statistical insight on the way languages are structured and used. The observation of language changes and interactions are also important in this process to determine how languages work. Another approach to a better understanding is through computer simulations. As

an example, we may cite the experiments developed by Steels (1998) "in which robotic agents and software agents are set up to originate language and meaning. The experiments test the hypothesis that mechanisms for generating complexity commonly found in bio-systems, in particular self-organization, co-evolution, and level formation, also may explain the spontaneous formation, adaptation, and growth in complexity of language". de Boer (2000) also develop similar research, where it is run a computer simulation of the emergence of vowel systems in a population of agents. "The agents (small computer programs that operate autonomously) are equipped with a realistic articulatory synthesizer, a model of human perception and the ability to imitate and learn sounds they hear. It is shown that due to the interactions between the agents and due to self-organization, realistic vowel repertoires emerge. This happens under a large number of different parameter settings and therefore seems to be a very robust phenomenon. The emerged vowel systems show remarkable similarities with the vowel systems found in human languages. It is argued that self-organization probably plays an important role in determining the vowel inventories of human languages and that innate predispositions are probably not necessary to explain the universal tendencies of human vowel systems". It seems a promising idea to use computer simulations of intelligent agents to show the role self-organization plays on the structuring effect observed on languages. Associated with statistical measures of the languages of the world, it is possible to compare the evidences and use the statistical information to build better simulations.

The ideas presented up to this point might be used in future research. There are a multitude of possible approaches and perspectives still to be analyzed, so it is not the intention to deplete all of them, but rather to have a wide view of the possibilities and interacting variables in order to create a better description and avoid possible contradictions with aspects not deeply studied.

Bibliography

Abe, S. and Suzuki, N. (2005). Scale-free statistics of time interval between successive earthquakes. *Physica A: Statistical Mechanics and its Applications*, 350(2-4):588–596.

Adamic, L. A. and Huberman, B. A. (2002). Zipf's law and the internet. *Glottometrics*, 3:143–150.

Aldrich, J. (1997). R. A. Fisher and the Making of Maximum Likelihood 1912-1922. *Statistical Science*, 12(3):162–176.

Alexander, L., Johnson, R., and Weiss, J. (1998). Exploring zipf's law. *Teaching Mathematics Applications*, 17(4):155–158.

Altmann, G. (1980). Prolegomena to Menzerath's law. *Glottometrika*, 2:1–10.

Altmann, G. and Arens, H. (1983). *Allgemeine Sprachwissenschaft, Sprachtypologie und Textlinguistik. Festschrift für Peter Hartmann*, chapter ,,Verborgene Ordnung" und das Menzerathsche Gesetz, pages 31–39. Narr, Tübingen.

Altmann, G., Schwibbe, M., and Kaumanns, W. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Olms.

Andrews, S. (1989). Frequency and neighbohood effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(5):802–814.

Apostel, L., Mandelbrot, B., and Morf, A. (1957). *Logique, Langage et Théorie de L'Information*. Presses Universitaires de France, Paris.

Baayen, R. H. (2001). *Word Frequency Distributions*. Kluwer, Dordrecht.

Baayen, R. H. (2003). *Probabilistic Linguistics*, chapter Probabilistic Approaches to Morphology. MIT Press.

Baddeley, A. D., Thomson, N., and Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, 14(6):575–589.

Bak, P. (1999). *How Nature Works: The Science of Self-organized Criticality*. Copernicus Series. Springer.

Balasubrahmanyana, V. and Naranan, S. (1996). Quantitative linguistics and complex system studies. *Journal of Quantitative Linguistics*, 3(3):177–228.

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., and Schoenemann, T. (2009). *Language as a Complex Adaptive System*, chapter Language Is a Complex Adaptive System: Position Paper, pages 1–26. Wiley-Blackwell.

Bloomfield, L. (1926). A set of postulates for the science of language. *Language 2*.

Bod, R., Hay, J., and Jannedy, S. (2003). *Probabilistic Linguistics*. MIT Press.

Bond, Z. S. and Garnes, S. (1980). *Perception and Production of Fluent Speech*, chapter Misperception of fluent speech. Erlbaum, N.J.

Bornstein, M. H. and Korda, N. O. (1984). Discrimination and matching within and between hues measured by reaction times: some implications for categorical perception and levels of information processing. *Psychological research*, 46(3):207–222.

Brent, M. R. and Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61:93–125.

Brown, G. D. A. and Neath, I. (2007). A temporal ratio model of memory. *Psychological Review*, 114(3):539–576.

Buk, S. and Rovenchak, A. A. (2007). Menzerath-Altmann law for syntactic structures in Ukrainian. *CoRR*.

Bybee, J. L. (2001). *Phonology and Language Use*. Cambridge University Press.

Bybee, J. L. (2003). *The evolution of language out of pre-language*, chapter Sequentiality as the basis of constituent structure, pages 109–132. John Benjamins.

Bybee, J. L. (2006). *Frequency of Use And the Organization of Language*. Oxford University Press.

Capek, M. J. (1983). Phoneme theory and umlaut: A note on the creation of knowledge. *Monatshefte*, 75(2):126–130.

Chen, S. F. and Goodman, J. (1998). An empirical study of smoothing techniques for language modeling. Technical report, Computer Science Group, Harvard University.

Chomsky, N. (1957). *Syntactic Structures*. Mouton.

Chomsky, N. (1968). *Language and Mind*. Harcourt, Brace and World, New York.

Chomsky, N. (1969). *Aspects of the Theory of Syntax*. The MIT Press.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*. Harper & Row.

Church, K. W. and Gale, W. A. (1991). A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language*, 5:19–54.

Cole, R. A. and Jakimik, J. (1980). *Perception and Production of Fluent Speech*, chapter A model of speech perception, pages 133–163. Lawrence Erlbaum Associates, Hillsdale, NJ.

Coleman, J. (2002). *Phonetics, phonology and cognition*, chapter Phonetic representations in the mental lexicon, pages 96–130. Oxford University Press.

Copelli, M. and Campos, P. R. A. (2007). Excitable scale free networks. *The European Physical Journal B: Condensed Matter and Complex Systems*, 56:273–278.

Coulmas, F. (2003). *Writing Systems: An introduction to their linguistic analysis*. Cambridge University Press.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley-Interscience, New York, USA.

Cristelli, M., Batty, M., and Pietronero, L. (2012). There is more than a Power law in Zipf. *Scientific Reports*, 2.

Dalattre, P. C., Liberman, A. M., and Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, 27:769–773.

de Boer, B. (2000). Self-organization in vowel systems. *Journal of Phonetics*, 28(4):441–465.

de Saussure, F. (1916). *Cours de linguistique générale*. Payot.

Deacon, T. (1997). *The Symbolic Species: The Co-Evolution of Language and the Brain*. Science: Linguistics. W.W. Norton.

Dresher, B. E. (2011). *The Blackwell Companion to Phonology*, volume 1, chapter The Phoneme, pages 241–266. John Wiley & Sons.

Düring, B., Matthes, D., and Toscani, G. (2008). A boltzmann-type approach to the formation of wealth distribution curves. Technical Report 08-05, Center of Finance and Econometrics, University of Konstanz.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, 24(2):143–188.

Etcoff, N. L. and Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition*, 44:227–240.

Everett, D. (2005). Cultural constraints on grammar and cognition in Pirahã: Another look at the design features of human language. *Current Anthropology*, 46:621–646.

Ferrer-i-Cancho, R. (2005a). The variation of Zipf's law in human language. *European Physical Journal B*, 44(2):249–257.

Ferrer-i-Cancho, R. (2005b). The variation of Zipf's law in human language. *European Physical Journal B*, 44(2):249–257.

Ferrer-i-Cancho, R. (2006). *Exact methods in the study of language and text. In honor of Gabriel Altmann*, chapter On the universality of Zipf's law for word frequencies, pages 131–140. Gruyter, Berlin.

Ferrer-i-Cancho, R. and Elvevåg, B. (2010). Random texts do not exhibit the real Zipf's law-like rank distribution. *PLoS ONE*, 5(3):e9411+.

Ferrer-i-Cancho, R. and Gavaldà, R. (2009). The frequency spectrum of finite samples from the intermittent silence process. *Journal of the American Society for Information Science and Technology*, 60(4):837–843.

Ferrer-i-Cancho, R. and Solé, R. V. (2002). Zipf's law and random texts. *Advances in Complex Systems*, 5(1):1–6.

Ferrer-i-Cancho, R. and Solé, R. V. (2003). Least effort and the origins of scaling in human language. *PNAS*, 100(3):788–791.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London*, 222:309–368.

Flemming, E. (2005). Deriving natural classes in phonology. *Lingua*, 115(3):287–309.

Frank, M. C., Everett, D. L., Fedorenko, E., and Gibson, E. (2008). Number as a cognitive technology: Evidence from Pirahã language and cognition. *Cognition*, 108(3).

Frisch, S. (1996). *Similarity and frequency in phonology*. PhD thesis, Northwestern University.

Fujimura, O. and Lovins, J. (1978). *Syllables and segments*, chapter Syllables as concatenative phonetic elements, pages 107–120. North Holland, Amsterdam.

Fujiwara, Y. (2004). Zipf law in firms bankruptcy. *Physica A: Statistical and Theoretical Physics*, 337(1-2):219–230.

Fukada, T. and Sagisaka, Y. (1997). Automatic generation of a pronunciation dictionary based on a pronunciation network. In *European Conference on Speech Communication and Technology (EuroSpeech 97)*, pages 2471–2474.

Gabaix, X. (1999). Zipf's law for cities: An explanation. *Quarterly Journal of Economics*, 114(3):739–67.

Gale, W. (1994). Good-Turing smoothing without tears. *Journal of Quantitative Linguistics*, 2.

Gale, W. A. and Church, K. W. (1994). What's wrong with adding one. In *Corpus-Based Research into Language. Rodolpi*.

Gale, W. A. and Sampson, G. (1995). Good-turing frequency estimation without tears. *Journal of Quantitative Linguistics*, 2:217–237.

Gernsbacher, M. A. (1994). *Handbook of Psycholinguistics*. Academic Press.

Givón, T. (1979). *On understanding grammar*. Academic Press.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264.

Good, J. (2008). *Linguistic Universals and Language Change*. Oxford University Press.

Grainger, J. (1990). Word Frequency and Neighborhood Frequency Effects in Lexical Decision and Naming. *Journal of Memory and Language*, 29:228–244.

Greenberg, J. H., Osgood, C. E., and Jenkins, J. J. (1966). *Universals of Language*, chapter Memorandum concerning language universals. M.I.T. Press.

Gregory, M. L., Raymond, W. D., Bell, A., Fosler-lussier, E., and Jurafsky, D. (1999). The effects of collocational strength and contextual predictability in lexical production.

Grignetti, M. (1964). A note on the entropy of words in printed english. *Information and Control*, 7:304–306.

Grzybek, P., Stadlober, E., and Kelih, E. (2006). The relationship of word length and sentence length: The inter-textual perspective. pages 611–618.

Hagège, C. (1986). *La Structure des langues*. Presses universitaires de France.

Hall, R. A. (1964). Review of 'lingua libera e liberta linguistica' by b. terracini. ., 40:288–291.

Hancock, J. T., Curry, L. E., Goorha, S., and Woodworth, M. T. (2004). Lies in conversation: An examination of deception using automated linguistic analysis. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, volume 26, pages 534–539, Mahwah, NJ: LEA.

Harris, C. M. and Wolpert, D. M. (1988). Signal-dependent noise determines motor planning. *Nature*, 394:780–784.

Hartley, R. V. L. (1928). Transmission of information. *Bell System Technical Journal*.

Herdan, G. (1966). *The Advanced Theory of Language as Choice and Chance*. Berlin.

Hockett, C. F. (1960). *A course in modern linguistics*. Macmillan.

Hopper, P. J. and Thompson, S. A. (1980). Transitivity in grammar and discourse. *Language*, 56(2):281–299.

Hopper, P. J. and Thompson, S. A. (1984). The discourse basis for lexical categories in Universal Grammar. *Language*, 60(4):703–752.

Hualde, J. I. (2004). Quasi-phonemic contrasts in spanish. In Chand, V., Kelleher, A., Rodríguez, A. J., and Schmeiser, B., editors, *WCCFL 23: Proceedings of the 23rd West Coast Conference on Formal Linguistics*, pages 374–398, Somerville, MA. Cascadilla Press.

Hurford, J. (1989). Biological evolution of the saussurean sign as a component of the language acquisition device. *Lingua*, 77(2):187–222.

Iverson, P. and Kuhl, P. K. (2000). Perceptual magnet and phoneme boundary effects in speech perception: Do they arise from a common mechanism? *Perception & Psychophysics*, 62(4):874–886.

Jakobson, R. and Waugh, L. R. (2002). *The Sound Shape of Language*. Mouton de Gruyter.

Jeffreys, H. (1939). *Theory of probability*. International series of monographs on physics. The Clarendon press.

Johnson, W. (1932). Probability: deductive and inductive problems. *Mind*, 41:421–423.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, 20:137–194.

Jurafsky, D. (2003). *Probabilistic Linguistics*, chapter Probabilistic Modeling in Psycholinguistics: Linguistic Comprehension and Production. MIT Press.

Jurafsky, D., Bell, A., Gregory, M., and Raymond, W. D. (2001). *In Frequency and the emergence of linguistic structure*, chapter Probabilistic relations between words: Evidence from reduction in lexical production, pages 229–254. John Benjamins, Amsterdam.

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall Series in Artificial Intelligence. Pearson Prentice Hall.

Kanter, I. and Kessler, D. A. (1995). Markov processes: Linguistics and Zipf's law. *Phys. Rev. Lett.*, 74(22):4559–4562.

Kay, P. and Regier, T. (2003). Resolving the question of color naming universals. *Proceedings of the National Academy of Sciences*, 100:9085–9089.

Kello, C. T., Brown, G. D., Ferrer-i-Cancho, R., Holden, J. G., Linkenkaer-Hansen, K., Rhodes, T., and Orden, G. C. V. (2010). Scaling laws in cognitive sciences. *Trends in Cognitive Sciences*, 14(5):223–232.

Klatt, D. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, 7:279–312.

Klatt, D. H. (1977). Review of the ARPA speech understanding project. *The Journal of the Acoustical Society of America*, 62(6):1345–1366.

Köhler, R. (1986). *Zur Linguistischen Synergetik: Struktur und Dynamik der Lexik*. Brockmeyer.

Köhler, R., Altmann, G., and Piotrovskiĭ, R. (2005). *Quantitative Linguistik /Quantitative Linguistics: Ein Internationales Handbuch /An International Handbook*. De Gruyter.

Kolguškin, A. N. (1960). *Linguistic and Engineering Studies in Automatic Language Translation of Scientific Russian Into English: Technical Report Phase II*. University of Washington Press.

Kruskal, J. B. and Wish, M. (1978). *Multidimensional Scaling*. Sage.

Laboissiere, R. (2010). preprint.

Labov, W., Ash, S., and Boberg, C. (2006). *The Atlas of North American English: Phonetics, Phonology, and Sound Change : a Multimedia Reference Tool*. Number v.1 in Topics in English Linguistic Series. Mouton De Gruyter.

Ladd, D. R. (2002). Distinctive phones in surface representation. In *8th Conference on Laboratory Phonology*, New Haven, Connecticut.

Ladefoged, P. and Maddieson, I. (1996). *The Sounds of the World's Languages*. Wiley-Blackwell.

Lamel, L. and Adda, G. (1996). On designing pronunciation lexicons for large vocabulary continuous speech recognition. In *ICSLP 96. Proceedings*, volume 1, pages 6–9.

Laplace, P. S. (1902). *A philosophical essay on probabilities*. John Wiley & Sons.

Leijenhorst, D. C. v., Weide, T. P. v. d., and Grootjen, F. (2005). A formal derivation of Heaps' law. *Information Sciences*, 170(2-4):263–272.

Li, W. (1992). Random texts exhibit zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, 38(6):1842–1845.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74:431–461.

Liberman, A. M. (1970). The grammars of speech and language. *Cognitive Psychology*, 1(4):301–323.

Liberman, A. M., Harris, K., Hoffman, H., and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54:358–368.

Lidstone, G. J. (1920). Note on the general case of the bayes-laplace formula for inductive or a posteriori probabilities. *Transactions of the Faculty of Actuaries*, 8:182–192.

Lü, L., Zhang, Z.-K., and Zhou, T. (2010). Zipf's law leads to Heaps' law: Analyzing their relation in finite-size systems. *PLoS ONE*, 5(12):e14139.

MacDonald, M. C. (1993). The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language*, 32:692–715.

Maddieson, I. (1984). *Patterns of Sounds*. Cambridge University Press.

Mandelbrot, B. (1954). Structure formelle des textes et communications. *Word*, 10:1–27.

Mandelbrot, B. B. (1953). An informational theory of the statistical structure of languages. In Jackson, B. W., editor, *Communication Theory*, pages 486–502.

Mandelbrot, B. B. (1965). Information theory and psycholinguistics. In Wolman, B. B. and Nagel, E. N., editors, *Scientific Psychology: Principles and Approaches*, pages 550–562. Basic Books Publishing.

Mandelbrot, B. B. (1982). *The Fractal geometry of Nature*. Freeman, New York.

Mandelbrot, B. B. (1999). *Multifractals and 1/f noise: wild self-affinity in physics (1963-1976)*. Selected works of Benoit B. Mandelbrot. Springer.

Manin, D. Y. (2008). Zipf's law and avoidance of excessive synonymy. *Cognitive Science: A Multidisciplinary Journal*, 32(7):1075–1098.

Manin, D. Y. (2009). Mandelbrot's model for Zipf's law: Can Mandelbrot's model explain Zipf's law for language? *Journal of Quantitative Linguistics*, 16(3):274–285.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Mit Press.

Martínez-Celdrán, E. (2004). Problems in the classification of approximants. *Journal of the International Phonetic Association*, 34:201–210.

Mendenhall, T. C. (1887). The characteristic curves of composition. *Science*, 9(214):237–249.

Mendenhall, T. C. (1901). A mechanical solution of a literary problem. *Popular Science*, 9:97–105.

Menzerath, P. (1928). über einige phonetische Probleme. In *Actes du premier Congres International de Linguistes*, pages 104–105, Sijthoff: Leiden.

Menzerath, P. (1954). *Die Architektonik des deutschen Wortschatzes*. F. Dümmler.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, T. G. B., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M., and Lieberman-Aiden, E. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331:176–182.

Mielke, J. (2005). Modeling distinctive feature emergence. In *West Coast Conference on Formal Linguistics XXIV*.

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

Miller, G. A. (1957). Some effects of intermittent silence. *The American Journal of Psychology*, 70(2):311–314.

Miller, G. A. (1965). *The Psycho-biology of language: an introduction to dynamic philology*, chapter Introduction. The MIT Press.

Miller, G. A. and Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America*, 27(2):338–352.

Miller, G. A. and Taylor, W. G. (1948). The perception of repeated bursts of noise. *The Journal of the Acoustical Society of America*.

Montemurro, M. A. and Zanette, D. H. (2011). Universal entropy of word ordering across linguistic families. *PLoS ONE*, 6(5):e19875.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47(1):90–100.

Nádas, A. (1985). On Turing's formula for word probabilities. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 33(6):1414–1416.

Nairne, J. S. (1988). A framework for interpreting recency effects in immediate serial recall. *Memory & Cognition*, 16:343–352.

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition*, 18:251–269.

Neath, I. and Nairne, J. S. (1995). Word-length effects in immediate memory: Overwriting trace decay theory. *Psychonomic Bulletin & Review*, 2(4):429–441.

Nowak, M. A., Komarova, N. L., and Niyogi, P. (2002). Computational and evolutionary aspects of language. *Nature*, 417:611–617.

Nowak, M. A., Plotkin, J. B., and Jansen, V. A. A. (2000). The evolution of syntactic communication. *Nature*, 404:495–498.

Odden, D. (2005). *Introducing Phonology*. Cambridge University Press.

Olson, D. (1994). *The World on Paper*. Cambridge University Press, Cambridge.

Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, 24(2):175–184.

Pierpont, W. G. (2002). *The Art and Skill of Radio-Telegraphy*. NOHFF.

Pierrehumbert, J. (1994). *Papers in laboratory phonology, vol. 3: Phonological structure and phonetic*, chapter Syllable structure and word structure: A study of triconsonantal clusters in English, pages 168–190. Cambridge University Press, Cambridge.

Pierrehumbert, J. B. (2003). *Probabilistic Linguistics*, chapter Probabilistic Phonology: Discrimination and Robustness. MIT Press.

Pinker, S. (2003). *The Language Instinct: The New Science of Language and Mind*. Penguin Books Limited.

Piotrovskii, R. G., Pashkovskii, V. E., and Piotrovskii, V. R. (1994). Psychiatric linguistics and automatic text processing. *Nauchno-Tekhnicheskaya Informatsiya, Seriya 2*, 28(11):21–25.

Pisoni, D. B. (1982). *Project SCAMP 1981: Acoustic Phonetics and Speech Modeling*, chapter In defense of segmental representations in speech processing. Institute for Defense Analyses, Communications Research Division, Princeton, NJ.

Pollack, I. (1952). The information of elementary auditory displays. *The Journal of the Acoustical Society of America*, 24(6):745–749.

Pombo, E. L. (2002). Visual construction of writing in the medieval book. *Diogenes*, 49(196):31–40.

Port, R. (2006). *Second Language Speech Learning: The Role of Language Experience in Speech Perception and Production*, chapter The graphical basis of phones and phonemes. John Benjamins, Amsterdam.

Port, R. (2007). How are words stored in memory? beyond phones and phonemes. *New Ideas in Psychology*, 25:143–170.

Port, R. and Leary, A. (2005). Against formal phonology. *Language*.

Pulvermüller, F. (2003). *The Neuroscience of Language: On Brain Circuits of Words and Serial Order*. Cambridge University Press.

Reetz, H. (2010). Simple UPSID interface. [online] `http://web.phonetik.uni-frankfurt.de/upsid.html`.

Saenger, P. (1991). *Literacy and Orality*, chapter The separation of words and the physiology of reading, pages 198–214. Cambridge University Press, Cambridge.

Saenger, P. H. (1997). *Space between words: the origins of silent reading*. Stanford University Press.

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996a). Statistical learning by 8-month-old infants. *Science*, 274:1926–1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., and Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, 70:27–52.

Saffran, J. R., Newport, E. L., and Aslin, R. N. (1996b). Word segmentation: The role of distributional cues. *Journal of memory and language*, 35:606–621.

Saffran, J. R. and Wilson, D. P. (2003). From syllables to syntax: Multi-level statistical learning by 12-month-old infants. *Infancy*, 4(2):273–284.

Samuelsson, C. (1996). Relating Turing's formula and Zipf's law. *CoRR*, cmp-lg/9606013.

Sapir, E. (1929). The status of linguistics as a science. *Language*, 5(4):207–214.

Schneier, B. (1996). *Applied cryptography: protocols, algorithms, and source code in C*. Wiley.

Sedelow, S. Y. and Sedelow, W. A. (1966). *The computer and literary style*, chapter A preface to computational stylistics, pages 1–13. Kent State University Press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*.

Shannon, C. E. (1951). Prediction and entropy of printed english. Technical Report 30, The Bell System Technical Journal.

Shoup, J. E. (1980). Phonological aspects of speech recognition. In Lea, W. A., editor, *Trends in Speech Recognition*, pages 125–138. Prentice Hall, Englewood Cliffs.

Steels, L. (1998). Synthesising the origins of language and meaning using co-evolution, self-organisation and level formation. pages 384–404. Edinburgh University Press.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, 64:153–181.

Studdert-Kennedy, M. (1976). *Contemporary issues in experimental phonetics*, chapter Speech perception, pages 243–293. Academic Press, New York.

Trubetzkoy, N. S. (1939). *Grundzüge der Phonologie*. Vandenhoeck und Ruprecht.

Tuldava, J. (1996). The frequency spectrum of text and vocabulary. *Journal of Quantitative Linguistics*, pages 38–50.

Tuldava, J. (2004). The development of statistical stylistics (a survey). *Journal of Quantitative Linguistics*, 11(1–2):141–151.

Vygotsky, L. S. (1934). *Thought and Language*. MIT Press, Cambridge, MA.

Wang, W. S.-Y., Ke, J., and Minett, J. W. (2004). Computational studies of language evolution. In Huang, C. and Lenders, W., editors, *Computational linguistics and Beyond*. Academia Sinica: Institute of Linguistics.

Weide, R. L. (2008). Carnegie Mellon Pronouncing Dictionary, release 0.7a. `http://www.speech.cs.cmu.edu/`.

Whaley, C. P. (1978). Word-nonword classification time. *Journal of Verbal Learning and Verbal Behavior*, 17:143–154.

Whorf, B. (1940/1956). *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*, chapter Science and linguistics, pages 207–219. MIT Press, Cambridge, MA.

Wickelgren, W. A. (1969). *Information Processing in the Nervous System*, chapter Context-sensitive coding in speech recognition, articulation, and development, pages 85–95. Springer-Verlag, New York.

Wilden, A. (2001). *System and Structure: Essays in Communication and Exchange*. International Behavioural and Social Sciences Classics from the Tavistock Press, 96. Routledge.

Yoshida, K., Watanabe, T., and Koga, S. (1989). Large vocabulary word recognition based on demi-syllable hidden Markov model using small amount of training data. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 1–4.

Young, G. and Householder, A. (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3:19–22.

Zipf, G. K. (1935). *The Psycho-biology of language: an introduction to dynamic philology.* The MIT Press.

Zipf, G. K. (1942). Children's speech. *Science*, 96:344–345.

Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology.* Hafner Pub. Co.

Ziv, J. and Lempel, A. (1978). Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536.