Language: universals, principles and origins.

Ramon Ferrer i Cancho

Barcelona, 2003.

Laboratori de Sistemes Complexos (GRIB−UPF)
&
Universitat Politecnica de Catalunya

Resum

En aquesta tesi s'investiguen vells i nous universals lingüístics, és a dir, propietats que obeeixen totes les llengües de la Terra. També s'estudien principis bàsics del llenguatge que prediuen universals lingüístics. En concret, dos principis referencials, mínim esforç de codificació i mínim esforç de decodificació, una reformulació dels principi de mínim esforç de G. K. Zipf pel qui parla i pel qui escolta. Els esmentats principis referencials prediuen la llei de Zipf, un universal de la freqüència de les paraules en el punt de màxima tensió entre necessitats de codificació i decodificació. Encara que s'han proposat processos trivials per explicar la llei de Zipf en contextos no lingüístics, aquí es recolza la significància d'aquesta llei per al llenguatge humà. Minimitzar la distància euclídea entre paraules sintàcticament relacionades dins frases és un principi que prediu projectivitat, un universal que afirma que els arcs entre paraules sintàcticament relacionades dins una frase no es creuen en general. D'una altra banda, aquesta minimització de la distancia física prediu (a) una distribució exponencial per a la distribució de la distància entre paraules sintàcticament relacionades (b) superioritat de l'ordre SVO en l'ús real de les llengües del món. Aquí es presenten propietats totalment noves de les xarxes de dependències sintàctiques, és a dir, distribucions de grau potencials, fenomen del món petit, *assortative mixing* i organització jeràrquica. Enlloc d'una gramàtica universal, es proposa una única classe d'universalitat per a les llengües del món. Sintaxi i referència simbòlica són unificades sota una única propietat topològica: connectivitat en la xarxa d'associacions senyal-objecte d'un sistema de comunicació. Assumint la llei de Zipf, no sols se segueix connectivitat sinó les propietats de xarxes sintàctiques reals esmentades més amunt. Per tant, (a) els principis referencials són els principis de la sintaxi i la referència simbòlica, (b) la sintaxi és el subproducte de principis simples de la comunicació i (c) les propietats esmentades de les xarxes de dependències sintàctiques han de ser universals si la llei de Zipf és universal, que és el cas. Es mostra que la transició a llenguatge és del tipus de les transicions de fase contínues en física. Per tant, la transició a llenguatge no va poder ser gradual. Es presenta el morfoespai reduït que resulta d'una combinació d'un principi de minimització de la distància i un principi de minimització de la densitat de connexions com una hipòtesi alternativa i una perspectiva prometedora per a xarxes lingüístiques que pateixin pressions per comunicació ràpida. La present tesi és única entre les teories sobre els orígens del llenguatge, en el sentit que (a) explica com les paraules o els senyals es combinen de forma natural per tal de formar missatges complexos, (b) valida les seves prediccions amb dades reals, (c) unifica sintaxi i referència simbòlica i usa ingredients que ja estan presents en els sistemes de comunicació animal, d'una forma que cap altra aproximació fa. El marc presentat és un canvi radical en la recerca dels universals del llenguatge i els seus orígens a través de la física dels fenòmens crítics. Els principis presentats aquí no són els principis del llenguatge humà, sinó els principis de la comunicació complexa. Per tant, els propdits principis suggereixen noves perspectives per a altres sistemes naturals de transmissió d'informació complexa.

Abstract

Here, old and new linguistic universals, i.e. properties obeyed by all languages on Earth are investigated. Basic principles of language predicting linguistic universals are also investigated. More precisely, two principles of reference, i.e. coding least effort and decoding least effort, a reformulation of G. K. Zipf's speaker and hearer least effort principles. Such referential principles predict Zipf's law, a universal of word frequencies, at the maximum tension between coding and decoding needs. Although trivial processes have been proposed for explaining Zipf's law in non-linguistic contexts, Zipf's law meaningfulness for human language is supported here. Minimizing the Euclidean distance between syntactically related words in sentences is a principle predicting projectivity, a universal stating that arcs between syntactically linked words in sentences generally do not cross. Besides, such a physical distance minimization successfully predicts (a) an exponential distribution for the distribution of the distance between syntactically related words and (b) subject-verb-object (SVO) order superiority in the actual use of world languages. Previously unreported non-trivial features of real syntactic dependency networks are presented here, i.e. scale-free degree distributions, small-world phenomenon, disassortative mixing and hierarchical organization. Instead of a universal grammar, a single universality class is proposed for world languages. Syntax and symbolic reference are unified under a single topological property, ie. connectedness in the network of signal-object associations of a communication system. Assuming Zipf's law, not only connectedness follows, but the above properties of real syntactic networks. Therefore, (a) referential principles are the principles of syntax and symbolic reference, (b) syntax is a by product of simple communication principles and (c) the above properties of syntactic dependency networks must be universal if Zipf's law is universal, which is the case. The transition to language is shown to be of the kind of a continuous phase transition in physics. Thereafter, the transition to human language could not have been gradual. The reduced network morphospace resulting from a combination of a network distance minimization principle and link density minimization principle is presented as an alternative hypothesis and a promising prospect for linguistic networks subject to fast communication pressures. The present thesis is unique among theories about the origins of language, in the sense that (a) it explains how words or signals naturally glue in order to form complex messages, (b) it validates its predictions with real data, (c) unifies syntax and symbolic reference and (d) uses ingredients already present in the animal communication systems, in a way no other approximations do. The framework presented is radical shift in the research of linguistic universals and its origins through the physics of critical phenomena. The principles presented here are not principles of human language, but principles of complex communication. Therefore, the such principles suggest new prospects for other information transmission systems in nature.

Language: universals, principles and origins.

Ramon Ferrer i Cancho

email: rferrer@imim.es

Complex Systems Lab (GRIB-UPF) and
Universitat Politècnica de Catalunya.

Final version

Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya

Llenguatge: universals, principis i orígens.

Programa d'intel.ligència artificial.
Intensificació de llenguatge.

Tutor de tesi
Dr. Horacio Rodríguez Hontoria
Departament de Llenguatges i sistemes informátics
Universitat Politècnica de Catalunya (UPC)

Director de tesi
Dr. Ricard V. Solé
Laboratori de sistemes complexos
Grup de recerca en informàtica biomèdica (GRIB)
Universitat Pompeu Fabra (UPF)

Acknowledgements

It is not easy to find words for expressing my great indebtedness to Ricard Solé. He taught me to think as a scientist. His teachings, projects and discussions about complex systems theory are behind the present inquiry. Many of the things he has taught me are beyond beyond the scope of science. He supported the crazy idea of a thesis about language when quantitative approaches are the issue of minority of researchers mostly in Central-East Europe and linguistics is dominated by purely axiomatico-deductive phylosophical approaches.

The Complex Systems Lab (and the old Complex Systems Research Group) has been a privileged scientific environment. Many members have been my teachers, scientific therapists, reviewers and friends: David Alonso, Javier G.P. Gamarra, Pau Fernández, José María Montoya, Romualdo Pastor, Jordi Delgado, Susanna Manrubia and Bartolo Luque. Pau's technical assistance deserves special mention.

I based Appendix B on an unpublished manuscript by Susanna Manrubia. Outlines of proofs and the underlying complete proofs in Chapter 7 are entirely due to Béla Bollobás and Oliver Riordan. Both have provided very hepful comments concerning other chapters. I thank Ludmila Uhlířová and Jan Králík for the opportunity to analyze the Czech corpus and the Dependency Grammar Annotator Project for their publicly available small Romanian corpus.

I would like to thank Toni Hernández for his deep and sharp reading of the present work. I am very grateful to the Santa Fe Institute and the Institute for Advanced Study (IAS) for his hospitality and stimulating environment. I want to thank Martin Nowak at the IAS and Reinhard Koëhler at Trier Universität for hospitality, supporting my research and helpful discussions. Reinhard Koëhler's synergetic understanding of language and independence with respected to mainstream theoretical linguistics have been of great value. I thank Béla Bollobás and Oliver Riordan for shedding light on the combinatorics of language.

I will try to list the remaining people whose support, comments or criticisms have influenced the present work (forgive me those I may have forgotten), i.e. Andrei G. Bashkirov, Dereck Bickerton, John Buck, Rafael Cases, Faust Diéguez, Josep Díez, Thomas Givón, Peter Harremoës, Dick Hudson, Natalia Komarova, David Krakauer, Alun Lloyd, Brenda McCowan, Francesc Martín, Alexander Mehler, Igor Melcŭk, George Miller, Partha Niyogi, Joshua Plotkin, Francesc Reina, Duane Rumbaugh, Ton Sales, Ryuji Suzuki, Jared Taglialatela, Flemming Topsøe, Enric Vallduví and Sandra L. Vehrencamp.

I thank Gerard Escudero and colleagues at the Universitat Politècnica de Catalunya for technical assistance. The research environment provided by the Grup de Recerca en informàtica biomèdica has been of great importance and so has been the staff at the UPF Mar Library. The present work could not have been taken place without the critical support of people beyond the scientific matters. Last statement goes to Marisa Casanovas, my parents, Araceli Gonzàlez, Anne Humes and Anna Jürgens. Marisa would have like to see my thesis written.

A man went to a circus in order to get a job.
The man who was in charge of evaluating the candidates asked him:
- What do you do?
- I imitate birds - replied the candidate.
- Oh, this is not interesting for us - said the evaluator.
And the candidate went away flying.

Eugenio, a Catalan stand-up comedian.

To Dorotea

Contents

Chapter 1

Introduction

Scientific work in general proceeds in three phases, the first observation, the second description of the observed phenomena, the third is the attempt to explain the results [...]. Most linguistic discussion is done in the domain of descriptions, which are constantly referred as theories *(e.g. Chomsky's standard and extended theories) probably in analogy to the purely formal axiomatic systems in logic and mathematics. The danger of this mistake lies in the fact that, as a consequence, description and explanation get confused. In fact, of the two, only explanation is not possible without a theory (in the proper sense, i.e. a system of laws and a number of additional prerequisites)*

Reinhard Köhler (1987, page 242)

Such methodology underlies the work presented here. A deep investigation about basic principles governing language has been carried out. Descriptions (i.e. statistical patterns) inspiring such principles are studied along with the predictions that such principles can make. The reduced set of principles studied here constitutes a simple but powerful theory whose predictions go beyond the statistical patterns they tentatively tried to explain. Moreover, such principles will help us to regard certain established principles in linguistics (such as projectivity) as consequences and not principles in the proper sense. More ambitious attempts of considering larger set of principles and their interrelations have been carried out (Köhler, 1986; Köhler, 1987), the so-called *synergetic linguistics*. The origins of a synergetic understanding of language can be traced back to G. K. Zipf studies (Prün, 1999). Here, special emphasis is made on the depth of the predictive power of every principle, with special regard to the origins and evolution of language. The principles studied here are not a consequence of a purely inductive methodology that was clearly the case of Greenberg's study about linguistic universals (Greenberg, 1968). Such principles are prolegomena to a theory of language.

The principles discussed here are the principles of whatever human language and individual speaker (whenever the individual has not been damaged), that is universal. It will be shown throughout the following chapters that such principles are capable of explaining linguistic universals, i.e. statistical patterns common to (almost) all languages.

Such a kind of research implies many challenges and leads to the formulation of a basic question. i.e. can we use the same principles for explaining communication systems ranging from the most simple communication systems such as the predator-type vervet monkey calls (Seyfarth, Cheney, and Marler, 1980a) to the most complex, e.g. the recursive syntactic systems in humans (Hauser, Chomsky, and Fitch, 2002)? There are two lines of research for that question. First, focusing on how different is human language from animal communication systems, which deceivingly leads to think human language is *off the chart* (Chomsky, 2002) and is a system for mental representation and thought more than a communication system (Chomsky, 1965a; Bickerton, 1990; Jackendoff, 1994). Second, investigating the conditions that could turn a simple communication system into a human-language-like system. Here we will take the second approach, benefiting from the essential differences stressed by the first line of research. We will show the basic ingredients for recursive syntactic systems stem from communication constraints (Chapters 7).

1.1 Some universal patterns

Formulating the principles of language and understanding the origins and evolution of language requires conveniently identifying its universals features. Some of them are examined in what follows according to the needs of the present work.

1.1.1 Zipf's law

The seek of the general principles governing communication systems implies the seek of statistical patterns. Tentatively, not all patterns are suitable for that aim. They must be universal (or almost universal), which is a wise mixture of how often the pattern appears and the diversity of conditions under which it appears. Let us illustrate it with a pattern that is the bulk of many chapters. $P(f)$, the proportion of words in a text whose frequency is f can be approximated by:

$$P(f) \sim f^{-\beta} \tag{1.1}$$

The previous equation is the so-called Zipf's law, that bears the name of George Kingsley Zipf, the linguist who made it popular (Zipf, 1932; Zipf, 1935; Zipf, 1942; Zipf, 1972a). If words in a sample text are ordered decreasingly by their frequency, the (normalized) frequency of a word is a power law of its rank (Zipf, 1972a), i, described in its simplest form as

$$P(i) \propto i^{-\alpha} \tag{1.2}$$

The first form (Eq. 1.1 is called the lexical spectrum (Tuldava, 1996) or the inverse Zipf's distribution (Cohen, Mantegna, and Havlin, 1997). Eq. 1.2 and

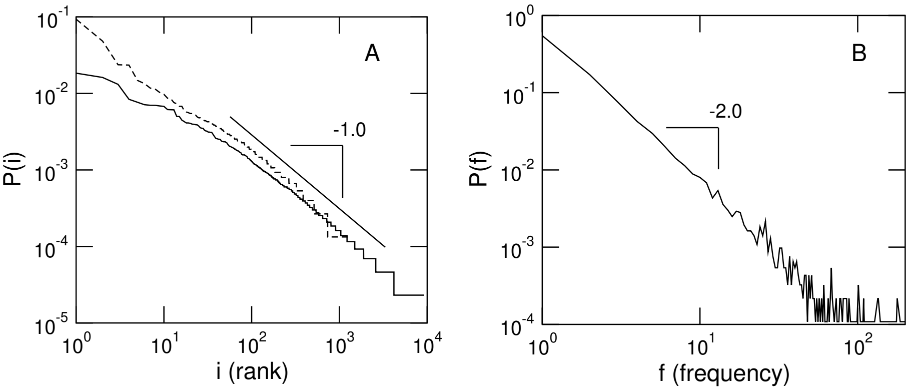

Figure 1.1: Frequency versus rank (A) and lexical spectrum (B) of Herman Melville's Moby Dick ($9,244$ different words). The dashed line in A shows the frequency versus rank for words having length 5, which is the average length of words in Melville's book (there are $1,248$ different 5-letter words). The exponents are (A) $\alpha \approx 1$ and (B) $\beta \approx 2 = \frac{1}{\alpha} + 1$ as expected.

1.1 are equivalent and their exponents obey (see Appendix B and for instance Naranan (1992; Naranan and Balasubrahmanyan (1992a))

$$\beta = \frac{1}{\alpha} + 1 \tag{1.3}$$

and their typical values are $\alpha \approx 1$ and $\beta \approx 2$. Although both the rank distribution and the word frequency spectrum can be modeled in many ways (Chitashvili and Baayen, 1993; Tuldava, 1996; Balasubrahmanyan and Naranan, 1996; Naranan and Balasubrahmanyan, 1998; Baayen, 2001), we adopt a plain power function for simplicity reasons. Unified representations of candidate Zipf's representations have been carried out (Zörnig and Altmann, 1995). Here, the term law refers to the strength of the empirical observation, that has been tested in different languages and authors (Balasubrahmanyan and Naranan, 1996). As far as we know, detailed and extensive study has only shown that the values of the exponents can vary from one sample to another (Balasubrahmanyan and Naranan, 1996) and more than one domain (Chapter 2 and Tuldava (1996), and Naranan and Balasubrahmanyan (1998)) is necessary for explaining the same sample. Fig. 1.1 shows the normalized frequency versus rank ($\alpha \approx 1$) and the lexical spectrum ($\beta \approx 2$) for Herman Melville's Moby Dick.

Such a regularity is known from the beginning of the XX century (Estoup, 1916; Dewey, 1923; Condon, 1928) and appears in oral and writing speech, in infants and adults (Zipf, 1942; McCowan, Doyle, and Hanser, 2002) and all languages where it has been tested (Balasubrahmanyan and Naranan, 1996). There seems to be no known exception to Eq. 1.1. Typically, $\beta \approx 2$ is found and rather exceptional values satisfy $1.5 \leq \beta \leq 3.4$ (Chapter 3). Zipf's law qualifies for a pattern inspiring general principles because of its ubiquity and its robustness. Given the huge amount of different distributions (Wimmer and Altmann, 1999) that could serve for arranging word frequencies, the question that has puzzled quantitative linguists over decades is *Why other distributions are not found?* Communication principles and the physics of critical phenomena give an answer in Chapters 3 and 4. Another reason for investigating Zipf's law is that it is assumed but not explained in recent models for the evolution of syntactic communication (Nowak, Plotkin, and Jansen, 2000; Nowak, 2000b). Is Zipf's law is meaningful (Chapter 5, it must be an ingredient for any theory of language evolution.

It is important to notice that sometimes Zipf's law refers to Eq. 1.1 with no specific value of β. Some other times, it precisely refers to $\beta \approx 2$. Sometimes, Zipf's law implicitly refers to words frequencies and sometimes to other linguistic units and systems. G. K. Zipf collected many rank-frequency relations (Zipf, 1972a). refer to it as the Zipf's law. In what follows, we will assume Zipf's law refers to the frequency distribution of units of reference (words in human language) in a generic communication framework. Since there are many exceptions to $\beta \approx 2$ even in a linguistic context, the exact value of the exponent will be explicitly mentioned when it is crucial.

1.1.2 Projectivity

The syntactic structure of a sentence can be usually specified by a network where arcs do not cross when drawn over words. This property is usually called projectivity in linguistic theory (Melčuk, 1989; Hudson, 1984). Arcs go from a modifier to its head, as in Fig. 1.2 A. The modifier is said to depend on the head. If the vertices of the sentence in Fig. 1.2 A are scrambled then multiple arc crossings (red circles) appear (Fig. 1.2 B). Fig. 1.2 B is far from the typical appearance of sentence structures. The majority of sentences in most languages are projective with the exception of particular cases (Melčuck, To appear). Projectivity qualifies thus as a linguistic universal. The origins of projectivity is a longstanding problem and an ultimate explanation has not yet been provided.

1.2 The principles

The principles presented here are based on the general assumption that communication has a cost. Communication implies different operations or processes whose cost communicating agents need to minimize. In other words, agents need to minimize the effort of such processes and operations. The principles presented here consider words as the basic unit. They are therefore lexical principles. Their validity for other linguistic units is beyond the scope of the present work. The principles are:

- Decoding least effort. Decoding, that is, the semantic interpretation of a word in a particular context has a cost. This principle states that such a

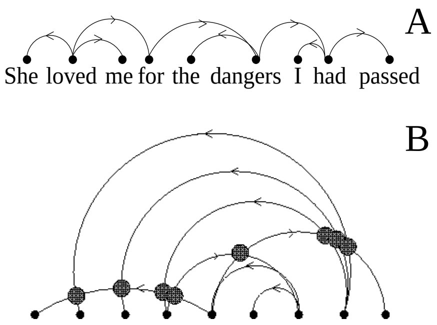

Figure 1.2: A. A sentence and its syntactic structure. B. The structure of the sentence in B is the same as that of A but the sequence of vertices is a random permutation of that of A. Gray circle indicate edge crossings. Links can only be drawn on the half plane formed by the straight line passing through the row of words.

cost must be minimized.

- Coding least effort. Coding, that is, finding the appropriate word for a certain meaning has a cost that is negatively correlated with the decoding least effort. Decoding least effort and coding least effort are opposite forces.

- Euclidean distance minimization. Sentences are a strings and have thus one dimension. The task of the speaker is to form such strings from mental representations in his brain. The task of the hearer is to map such strings with its mental representations. The Euclidean distance minimization principle states that the distance between syntactically dependent words in the same sentence must be minimized.

- Network distance minimization. Fast navigation in linguistic networks requires the distance between vertices is minimized. Syntactic dependency networks are an example of linguistic network where vertices are words and arcs are syntactic dependencies.

- Link density minimization. Links in linguist networks have a cost that must be minimized. The distance between vertices is generally negatively correlated with the amount of links used. Networks density and link density minimization are opposite forces.

Resorting to some of the principles will allow explaining the universal patters presented in Section 1.1 along with making further predictions. Decoding least effort is a revisit of Shannon's communication theory. Decoding and coding least effort are considered in system theoretical linguistics (Köhler, 1986; Köhler, 1987) but the predictions made here are totally new. Coding and decoding here refers to the mapping between words and meanings. Referential coding and referential decoding will not be used for brevity reasons. The remaining principles are a novel contribution of the present work. The network distance minimization principle and the links density minimization principles are the most hypothetical among the five and require future exploration beyond the scope of the present work.

Such principles can be classified into two classes, i.e. referential principles communication and network principles.

1.2.1 Referential principles

Different prominent linguists have emphasized that human language is much more than a mere communication system (Hauser, Chomsky, and Fitch, 2002; Bickerton, 1990; Jackendoff, 1994). The existence of a universal grammar with its own set of principles as been hypothesized (Uriagereka, 1998). If the universal grammar is said to be uniquely human, what can their principles tell us about simpler communication systems and how such systems can reach higher levels of complexity? Probably nothing, because universal grammar is strictly tied to syntax and non-human species seem not to have syntax. Nonetheless, the gap between simple communication and syntax will be bridged here using a novel approach.

Communication undergoes many constraints and pressures. Information theory provides a basic scheme that helps to understand the goals and constraints of communication. Under that view, communication takes place between a sender (e.g. the speaker) and a receiver (e.g. the hearer) (Fig. 1.3). The task of the sender is to code a message that the receiver has to decode. The goal of communication is that the receiver interprets the message intended by the sender. General representations of the basic scheme include a source of noise (dashed box in Fig. 1.3) making that the receiver misunderstands the code delivered by the sender.

Shannon was the first to mathematically formalized the goal of communication (Shannon, 1948). He defined a measure, the rate of information transmitted (transinformation), a communication system has to maximize. Assuming, S, a set of signals (i.e. codes) and R, a set of objects (i.e. messages) the transinformation measure for a biological communication is defined as

$$I(R,S) = H(R) - H(R|S)$$

where $H(R)$ is the entropy associated to object frequencies and $H(R|S)$ is the average entropy associated to the interpretation of every signal (see (Ash, 1965) for further details). Hereafter, an agreement between sender and receiver about the possible interpretations of a code will be assumed (see Section 10.3 for a discussion of this assumption). Thus, it can be intuitively seen that the higher the amount of meanings of a word, the higher the possibility of misinterpreting that word. The goal effective communication only consists of minimizing the decoding communication function we define (assuming $I(R,S)$ is normalized)

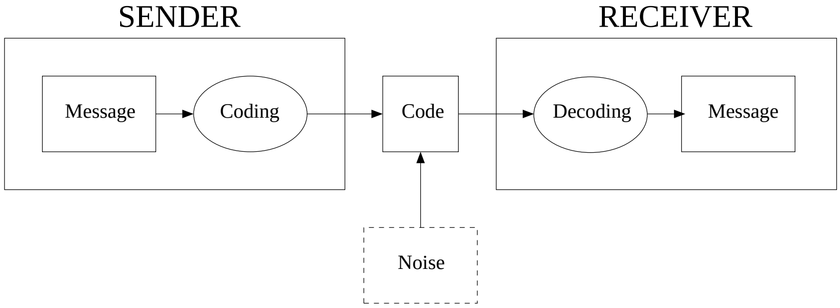

Figure 1.3: Basic scheme of a communication system.

as

$$E_D = 1 - I(R, S) \tag{1.4}$$

which is equivalent to minimizing

$$E_D = H(R|S) \tag{1.5}$$

if coding warrants $H(R)$ is constant (Chapter 3 makes use of this assumption). At a time where artificial communication devices where flourishing, Shannon's concerns about communication stop here.

We will say that E_D measures the decoding effort. It will be shown that the decoding effort does not lead to Zipf's law with $\beta \approx 2$ (Chapter 3). Instead, it will lead to all signals having a similar frequency (provided that objects have a similar frequency; see Chapter 5). One of the major findings presented here is that Zipf's law (with $\beta \approx 2$) results from taking into account, E_C, the encoding effort. Encoding effort has been never been considered in the evolution of language and artificial intelligence literature and consistently the natural emergence of Zipf's law has never been reported in those works. Psychological constraints limit the availability of words in humans. The lower the frequency of a word, the lower its availability, the so-called word frequency effect. Getting the appropriate words has a cost. Accordingly, we define the dual communication function as

$$\Omega = (1 - \lambda)E_C + \lambda E_D \tag{1.6}$$

where $E_C = H(S)$ measures the coding effort and $0 \leq \lambda \leq 1$. Notice that Ω is a generalization of Shannon's information transfer provided $H(R)$ is constant. Shannon's framework is recovered for $\lambda = 1$. Another novel contribution of the work presented here is to take into account the positive correlation between frequency and number of meanings (Reder, Anderson, and Bjork, 1974; Köhler, 1986; Manning and Schütze, 1999). so that the word frequency effect becomes the word meaning effect: *the more meanings a word has, the higher its availability.* By doing so, Zipf's law (with $\beta \approx 2$) emerges at the maximum tension between encoding and decoding effort (Chapter 3), that is $\lambda \approx 1/2$. When such a dual satisfaction of encoding and decoding needs is not present, exponents other $\beta \approx 2$ or even non-power distributions should be expected. Interestingly, Chapter 4 shows scaling is preserved even when only E_D is prescribed, consistent with variations of $\beta \approx 2$ in human language.

The models presented here assume that there is no source of noise. Nonetheless, the coding least effort principle can be regarded as a source of noise (i.e. misinterpretation) operating inside the sender at the coding stage Fig. 1.3. Mathematical approaches to the evolution of language assume similarities between codes are the source of noise (Nowak and Krakauer, 1999; Nowak, Krakauer, and Dress, 1999; Nowak, Plotkin, and Krakauer, 1999; Nowak, Plotkin, and Jansen, 2000; Nowak, 2000b). The standard communication scheme (Shannon (1948),Ash (1965); Fig. 1.3) and mathematical approaches to the evolution of language assume noise is external to both sender and receiver.

G. K. Zipf hypothesized that the law bearing his name was due to a tension between unification (one word with multiple meanings) and diversification forces (distinctly different words for different meanings), a principle of least effort in his own words. The validity of G. K. Zipf hypothesis was never shown. An information theory approach in Chapter 3 and 4 inspired in his hypothesis will explain Zipf's law. Nonetheless, it has to be noted that if the effort for the hearer, the decoding effort, is defined as the vocabulary size the maximum tension between hearer and speaker needs disappears and Zipf's law does not emerge (see Chapter 3 for further details). Besides, just minimizing the amount of meanings per word does not lead to Zipf's law, although it is a way of minimizing the decoding effort (Chapter 4). G. K. Zipf does not draw a clear distinction between the the relationship between vocabulary size minimization and vocabulary versatility maximization. Therefore, Zipf's principle of least effort must be seen as a rough intuition. It is important to keep in mind that talking about vocabulary size minimization in the context Zipf's law is, in the best case, a side-effect of a more complex definition of the coding effort. Later interpretations (Ball, 2003) can be deceiving when equating coding least effort with vocabulary size. Distinguishing between causes and consequences is necessary.

A huge amount of different mechanisms reproducing Zipf's law have been proposed. During a first wave at the beginning of the first half of the XX century, G. Miller, H. A. Simon and B. Mandelbrot proposed different explanations. H. A. Simon proposed a multiplicative stochastic process of the type *rich-gets-richer* (Simon, 1955) (a birth process without death). G. Miller (Miller, 1957) and B. Mandelbrot (Mandelbrot, 1966) present intermittent silence, a process concatenating characters chosen at random. The characters set includes spaces behaving as word separators. Words are defined as maximal contiguous sequences of letters without spaces. N. Chomsky and G. Miller revisited intermittent models in a book (Miller and Chomsky, 1963) that is considered as a reference against Zipf's law meaningfulness. B. Mandelbrot proposed word length minimization (Mandelbrot, 1953). After such initial wave many models and explanations have been proposed (Chapter 5), including a rediscovery of the intermittent silence models by W. Li (Li, 1992). Therefore, Zipf's law can be explained in multiple ways. A careful look reveals that many explanations or null hypothesis make assumptions making no sense in a linguistic context. For instance, W. Li, B. Mandelbrot and N. Chomsky assumptions for intermittent silence model assume that words are not chosen from a finite size repository (the so-called mental lexicon) and that words are used according to their meaning. In fact, no model known takes into account that words have meaning, except for the models presented here (Chapter 3 and 4). Models where word meaning reproduce Zipf's law but do not qualify as an explanation for Zipf's law. The models presented here are not the ultimate answer for Zipf's law (which would be groundless from a Popperian perspective (Popper, 1968)), but contain an essential ingredient that is forgotten by all existent models, i.e. reference. Chapter 5 contains a critical review of existent explanations or null hypothesis.

1.2.2 Network principles

If a communication system uses strings of words for coding messages as humans do, the cost of a syntactic links is positively correlated with the distance between linked words (Ueno and Polinsky, 2002; Gibson, 2000; Hawkins, 1994). Speakers and hearers must minimize the distance between linked words which is consistent with the fact that most of syntactic links take place at very short distances (Chapter 8). Minimizing the Euclidean distance between linked words, inspired in the minimum linear arrangement problem (Díaz, Petit, and Serna, 2002), will bear such responsibility and explain projectivity (Chapter 8).

Fast communication pressure is obvious in the usual agonistic context where animal communication takes place (Hauser and Nelson, 1991) and it is hypothesized as a driving force in the evolution of human language (Lieberman, 1991a). Evidence of fast communication pressures are found at different levels of human language. The sounds of human speech allow us to transmit phonetic segments at an extremely rapid rate, up to 20 segments per second. In contrast, human beings can not identify non-speech sounds at rates that exceed seven to nine items per second (Lieberman, 1992). The coding least effort is a consequence of fast communication pressures, since the time needed for getting a word is positively correlated with the coding effort.

Fast communication is also a pressure for mental navigation in different types of linguistic networks. Different experiments in psycholinguistics show that words are interconnected in different ways. It is known that the network topology is crucial for the speed at which navigation can be performed (Watts and Strogatz, 1998). d the average vertex-vertex distance is an inverse measure of the navigation speed. We will define the network distance least effort principle as minimizing d. If links have a cost, there is a conflict between ρ network density and minimizing d. We define the link density minimization principle as minimizing ρ. The network with the minimum d is a complete graph, but it is the most expensive topology. Such a conflict is described by an energy equation combining the conflicting needs through a parameter λ

$$\Omega(\lambda) = \lambda d + (1 - \lambda)\rho$$

The previous energy function provides a network morphospace with only five network types (Chapter 9).

1.3 The evolution of language

Scholars have made an effort to outline the differences between human language and other forms of communication in other species, in a way that, language, a complex form of communication, refers uniquely to humans and communication to the remaining species. Scholars argue the hallmark of human language is syntax or symbolic reference. The core of syntax is the ability of combining elements from a finite set (e.g. words) and yielding a potentially infinite array of discrete expressions, the so-called discrete infinity (Hauser, Chomsky, and Fitch, 2002). Symbolic reference is the highest form of reference, that is the mapping between signifiers (e.g. sounds) to objects of reference (e.g. meanings). Three increasing levels of reference are distinguished: iconic, indexical and symbolic. Briefly, reference is iconic when the mapping between signifiers and objects is made trough physical similarity. Reference is indexical when the mapping between signifiers and objects is made trough temporal or spatial correlation. Convention is the way signifiers and objects get linked in symbolic reference. Besides, symbolic reference implies interrelations between signifiers that are not present in lower forms of reference (Deacon, 1997).

Scholars are divided into proponents of syntax (Hauser, Chomsky, and Fitch, 2002) or symbolic reference (Deacon, 1997; Donald, 1991; Donald, 1998) as the crux of human language. Besides the present work (Chapter 7) there is no conciliating and integrative approach.

Over the last centuries, scholars have tried to provide an answer for the following question concerning the evolution of human language:

- Is human language unique?
- If there a gap between human language and the communication systems of other species?
- Did (human) language appeared gradually or suddenly on Earth?
- Is human language the result of the evolution (extension) of a simple communication system or a by-product of another function?

The following sections explore the previous questions.

From the one hand, human language may have evolved by extension of preexisting communication systems or exapted away from other functions (e.g. spatial or numerical reasoning, Machiavellian social scheming, tool-making) (Hauser, Chomsky, and Fitch, 2002). From the other hand, assuming language is an extension of simple communication systems, one has to propose what the driving mechanism is. Darwinian evolution (Nowak and Krakauer, 1999; Nowak, Krakauer, and Dress, 1999; Nowak, Plotkin, and Krakauer, 1999; Nowak, Plotkin, and Jansen, 2000; Nowak, 2000b) or learning constraints (Kirby, 2000; Kirby, 2002a) are a possible explanations. If evolution has to choose among syntactic and non-syntactic communication, just formalizing the conditions under which syntactic communication is selected does not provide an ultimate answer, since it does not explain why words should naturally combine. Similarly, how can syntax be selected by evolution when there is no syntax at all? If syntax is assumed and natural selection has to make a decision about selecting it or not, it is not a totally fair game, since how syntax naturally appears is not explained. If learning constraints choose syntactic communication without the help of Darwinian selection, providing the system with a phrase structure grammar is not a fair game since such a formalism implies recursion. Accordingly, we argue here that natural word combinations must be a by-product. Darwinian selection can not operate directly on syntax but also on another function. The contribution of this thesis consist of an explanation where such function is non-syntactic communication itself, and not a totally different function. More precisely, we will show natural word combination results from a solving a conflict between coding and decoding least effort principles (Chapter 7).

1.3.1 Is human language actually unique?

The language-communication distinction is based on the following fact: there is no positive evidence of other species having a communication system as ours. The conclusion requires explanation. There are species for which studies of their communication systems have been carried out. Studying the communication system of a species (if there is any) is generally a very complicated task. We will simplify such a task by means of two basic necessary (but not sufficient) conditions for language: reference and combinatorics. Combining simple units seems a condition close to infinite discreteness. Many species produce songs based on the combinations of units. Singing birds (Gardner, Cecchi, and Magnasco, 2001) are clear examples that combinatorics does not imply reference. Nonetheless, when the higher species are studied, higher attention is paid. When whales sing (Noad et al., 2000) or gibbons sing or produce long distance calls (Geissmann, 2000; Hauser, 1996), their vocalizations must be more conspicuously analyzed. Unfortunately, no conclusive evidence of reference has been reported. In some lucky cases, researchers conclude the vocalizations of a species have clear mappings with external stimuli. In that cases, evidence of combinatorics is not generally found although they are usually concentrated on a limited range of utterances (Davison, 1997) (e.g utterances that are relatively easy to observe and to interpret). Table 1.1 summarizes the communication systems of different systems according to our current knowledge. Some studies have shown species combining units that may carry meaning. Nevertheless, whether or not combinations of these units into sequences encode something more than the concatenation of separate meanings of the units is a difficult question that has not yet been answered (Ficken, Hailman, and Hailman, 1994).

No negative conclusions have reached for the utterances of certain species. Many cetaceans fall into this category (dolphins, whales, belugas,...). When

		Reference	
		No	Yes
Combinatorics	No		Vervet monkeys (Seyfarth, Cheney, and Marler, 1980b)
	Yes	Whales (Noad et al., 2000), Gibbons (Hauser, 1996; Geissmann, 2000), singing birds (Gardner, Cecchi, and Magnasco, 2001)	Humans

Table 1.1: Classification of communications systems according to the presence or absence of combinatorics and reference.

scholars make equivalent language and humans, they are neglecting the latter dark cases (Chomsky, 1972; Chomsky, 1988a; Hauser, 1996). Species producing complex signals that can hardly be broken into units or the reference is impossible to assess. Recently, dolphin vocalizations have been systematically segmented and categorized using automatic techniques (McCowan, Hanser, and Doyle, 1999; McCowan, Doyle, and Hanser, 2002). Although Zipf's law is found, there is no ultimate answer about the suitability of the segmentation and categorization techniques used (Janik, 1999). Reference is suspected for that species (Janik and Slater, 1998; Janik, 2000), although it not been proved. Here it is shown that Zipf's law plus reference allow making relevant predictions about the possibility of combinatorics (Chapter 7).

Human language uniqueness is rooted in an anthropocentric understanding of nature. Our anthropomorphic bias limits the way we look at non-human species communication systems (Savage-Rumbaugh, 1999). We have failed in decoding the complex utterances of different species, but we believe we are unique. Our way of exploring other species complex signals relies on two anthropocentric implications

1. If we can not understand it then we assume there is no language.

2. If there is no reference then language can not exist.

The first implication reflects our inability to deal and integrate doubts into our cosmovision (*I do not know* is a bad answer). Let us show the second implication is false with the following mental experiment. Imagine an extraterrestrial intelligent organism wants to determine whether we humans have language. Imagine the extraterrestrial researcher has a technique for breaking our complex vocalizations into words. Such researcher wants to determine first if words are mostly referential. Therefore, he decides (let us say 'he' and not 'it' assuming it has language) to choose the most frequent words for their study because they capture most of humans' behaviour. Imagine for that study, they choose the seven most frequent words in English. The words are *the, of, and, to, a, in, that.* After detailed study, the extraterrestrial researcher concludes, as a human would agree, that the seven most frequent words have no referential power. The second implication leads to the following conclusion *humans have no language*, which is a contradiction. The lesson is the following: starting with the simplest hypothesis with non-human species can fail. In some cases the most convenient hypothesis is *language* and not *simple communication.*

Information coding and decoding systems pervade nature. Human speakers code words from meanings that hearers must guess. Communication on the fly, that is, communication requiring coder and decoder cooccur in time and space is present in multiple forms. Human oral language is an example among many. Sound is not, by far, the only means, not only in humans, but in other cases. Vervet monkeys alarm calls containing information about the type of threatening predator (Seyfarth, Cheney, and Marler, 1980b) are simple examples. Nature stored information in genes millions of years before humans invented writing systems. Genes code for proteins using intermediate RNA. Even very simple organisms, such as bacteria, communicate exchanging genetic material by contact (Losick and Kaiser, 1997; Miller, 1998). Bacteria develop drug resistance by means of genetic communication. Honey bees dance for indicating with high precision the position of distant sources of food (Frisch, 1967). We humans may be unique using our definition of language, but we are not by far the first species to use non-trivial information transmission systems.

1.3.2 No gap but abrupt emergence

There has been a long debate about the nature of human language. Some approaches claim that there must have been a *continuous* or *gradual* adaptation from animal communication systems to human language (Lieberman and Kosslyn, 2002; Pinker and Bloom, 1990; Newmeyer, 1991; Brandon and Hornstein, 1986; Corballis, 1991; King, 1994; Bickerton, 1981). Some other approaches argue there must have been a *discontinuity* between the two stages (Chomsky, 1988a; Chomsky, 1991; Bickerton, 1990; Bickerton, 1996). Is there a possibility to reconcile such views? Initially, the former are marvelled by the striking similarities between human and some other species (the striking similarities between humans, great apes and cetaceans social behavior, learning and communicative skills) and the power of Darwinian selection. The latter make emphasis in the difference between language and the remaining forms of communication known. The hope for reconciliation comes from the fact that, generally and to some extent, the opposite hypothesis come from different perspectives and not from conflicting argumentations inside the same framework. Taking the eye of physics, we will show that an integrative position between both views is a sound candidate for explaining how human language actually emerged (Chapter 3,7).

Proponents of discontinuity argue that human language is thousands of miles away from (poorly known) animal communication systems. Having infinite discreteness and not having it makes a radical difference. One of the greatest successes of Noam Chomsky's philosophical inquiry is that such a change can not be gradual. It is important to notice that not gradual and discontinuous are not equivalent, at least in the physics of critical phenomena. The physics of critical phenomena calls phase transitions to radical changes between qualitatively different phases (Binney et al., 1992). The transition from boiling water to vapour is a phase transition where liquid water and vapour are different phases. Interestingly, two types of phase transitions are basically distinguished, e.g. continuous (second order) and discontinuous (first order). Most of popular phase transitions such as evaporation (in regular conditions) and melting are discontinuous. The crucial difference is that continuous phase transitions exhibit intermediate configurations between phases. The appearance of spontaneous magnetization in a ferromagnet (like iron), if it is cooled below a certain critical temperature, is continuous. Discontinuity is the word used in the evolution of language for emphasizing the distance between language and communication (let me avoid the equivalence pairs language-humans and communication animals) and there it is sometimes used without clear convention about whether it implies intermediate stages (that should be rare) or no intermediate stages at all.

We will show (Section 1.3.3 and Chapter 7) that the transition to syntax should have been a continuous phase transition, which implies the possibility (at least theoretically) of existent species with intermediate stages between language and simple communication. The transition between one phase and another is governed by a control parameter. A threshold value of such a parameter determines the end of one phase and the beginning of the other. Intermediate configurations exist in a narrow domain of a continuous phase transition around the threshold value. Therefore, although we argue intermediate situations may exist, they should be rare.

1.3.3 The transition to syntax must be continuous but sharp

Human syntax can be modeled in a simple way with a network in which words are vertices and links are syntactic relationships between pairs of words. Tentatively, linking words according to such a graph will lead to syntactically well formed sentences if word relative order is not taken into account. This is the basic approach of dependency syntax formalism (Melčuk, 1989; Hays, 1964; Sleator and Temperley, 1993) where the structure of a sentence itself is described in terms of a network. Our model must be regarded as a the skeleton for syntax and not as syntax in the strict sense because a real well-formed sentence often requires word order to be satisfied (Sleator and Temperley, 1991) and other details such as link direction (Melčuk, 1989) and form agreement between a head word and its modifier (Akmajian, 1995). The crucial difference is that such sort of grammar does not imply (but allows) recursion (Hauser, Chomsky, and Fitch, 2002). Nonetheless, such a network clearly contains our major concern, that is, that some words are glued together. We assume we have

n vertices and that every pair of vertices is linked with probability p. We will refer to such a random network as $G_{n,p}$ (Bollobás, 2001). A minimal well-formed phrase linking u and v will be made up of u,v and the vertices in whatever path between u and v. We will say that a set of vertices V is a protosyntactic system when it will be possible to build a syntactically well formed phrase for whatever pair of vertices (u, v) in V. The number of different phrases that can be generated by a combinatorially powerful set of word with n vertices is greater than $n(n-1)/2 \in O(n^2)$. If such phrases are integrated into sentences, the number of messages that can be constructed is theoretically unlimited. We also define the syntactic power of $G_{n,p}$ as the maximal set of vertices that is a protosyntactic system. We define Σ as the number of vertices in the largest connected component (Bollobás, 1998).We will say $G_{n,p}$ is a full protosyntactic system when it is connected (i.e. $\Sigma = n$). When $p = 0$, our model $G_{n,p}$ is clearly in the single word phase we find in non-human animals and children at the early stages (Johnson, Davis, and Macken, 1999). We call p^* the smallest value of p warranting $G_{n,p}$ is connected. The transition from a non-syntactic phase to a protosyntactic phase can be mapped into a transition from unconnectedness to connectedness in $G_{n,p}$. It is known that the transition will be sharp (Bollobás, 2001).

This supports the intuition among certain linguists that the evolutionary transition to syntactic communication would have been a sort of *big bang* (Newmeyer, 2000; Chomsky, 1988a; Chomsky, 1991; Bickerton, 1990; Bickerton, 1996). Such a simple model explains the gap between animal and human communications, since close values of p may radically differ in their protosyntactic power. The model also explains why the transitions to syntactic communication in children should be fast after a period of rather short period of one or two words utterances (Johnson, Davis, and Macken, 1999). Nevertheless our answer is incomplete. A transition to connectedness or near connectedness is not an explanation for *why* syntax emerged but a partial explanation for *how* it emerges. From the one hand, we need to explain *why* words 'glue'. Words 'glue' through their meaning(s) is the assumption defended here (Chapter 7).

Here we will (Chapter 3) show that Zipf's law (with $\beta \approx 2$) and the corresponding optimal satisfaction of hearer and speaker needs is in fact an intermediate configuration between a no communication phase (one word for all meanings) and a perfect communication phase (a one-to-one map between words and meanings). We will show the corresponding network of signal-signal associations (e.g. word-word associations), a rudimentary form of syntax and symbolic reference, has a degree distribution with a power tail whose exponent is $\gamma \approx 2$ (Chapter 7). $\gamma \approx 2$ is the threshold for a continuous phase transition from a disconnected to a connected phase in networks with a pure power degree distribution (Newman, Strogatz, and Watts, 2001), suggesting syntactic interactions operate at the point where the possibility of linking whatever pair of words holds trivially.

Principle	Prediction
Coding and decoding least effort (Chapter 3,7)	$p_k \sim k^{-2}$
	$p_f \sim f^{-2}$
	Maximum entropy (MaxEnt) in p_k
	Syntax
	Symbolic reference
Decoding least effort + MaxEnt (Chapter 4)	$p_k \sim k^{-\beta}$
	$p_f \sim f^{-\beta}$
Euclidean distance minimization (Chapter 8)	$p_d \sim e^{-cd}$
	Projectivity
	SVO order superiority
	SVO order in creoles
	SVO order for the first language spoken on Earth
Network distance minimization + link density minimization (Chapter 9)	$q_k \sim e^{-ck}$
	$q_k \sim k^{-\gamma}$
	star graph
Network distance minimization (Chapter 9)	Complete graph
Link density minimization (Chapter 9)	Poissonian graph

Table 1.2: Principles and their predictions. p_f is the proportion of signals (e.g. words) whose frequency is f, p_k is the proportion of signals with k connections with objects (of reference), p_d is the proportion of syntactically linked words at distance d in a sentence, q_k is the proportion of words with k connections.

1.4 Summary

Table 1.2 summarizes the principles presented above along with their predictions. Predictions mostly take the form of statistical patterns. Some statistical patterns are already known and some of them are novelly presented here. Such statistical patterns qualify as linguistic universals, i.e. features common to (almost) all languages.

The more often it appears, Different trends are followed in linguistics. Typology is devoted to compare languages and find common and diverging traits. The common traits and the rules explaining differences and classes of languages are the so-called linguistic universals. Quantitative linguists is devoted to find statistical patterns that are often not language dependent and are extremely universal.

The remaining chapters are organized as follows. Chapter 2 shows the existence of a core lexicon where Zipf's law (with $\beta \approx 2$) is found. The core lexicon will be the implicit domain of the following chapters. Chapter 3 shows that Zipf's law with $\beta \approx 2$ emerges from the maximum tension between coding and decoding least effort principles and presents a novel scenario for the origins of language. Chapter 4 maps β with the decoding effort and makes a divides communication systems into simple and complex depending on the strategy used for minimizing the decoding effort. Chapter 5 is critical review of existent hypothesis for Zipf's law and provides novel arguments for Zipf's law meaningfulness. Chapter 6 presents a totally new approach to syntactic universals. Chapter 7 explains how a rudimentary form of syntax and symbolic reference follows from Zipf's law. Real syntactic patterns found in Chapter 7 are just a consequence of Zipf's law. Chapter 8 shows that the locality of syntactic dependencies within sentence as well as projectivity are a consequence of an Euclidean distance minimization principle. Such a principle predicts the circumstance under which SVO order should appear. Chapter 9 presents a network distance minimization principle as an unsuccessful alternative for the Euclidean distance minimization principle and as suggestive prospect for linguistic networks subject to fast communication pressures. Chapter 10 is a comparative analysis of the present work and other views in linguistics and the evolution of language. Chapter 11 summarizes the major contributions presented in the previous chapters emphasizing the original aspects of the present work. See Appendix G for the publications upon which the different chapter are based.

Chapter 2

A core and a peripheral lexicon

2.1 Introduction

It can be observed in the plots of (Zipf, 1972a; Casti, 1995; Tsonis, Schultz, and Tsonis, 1997) that Zipf's law provides a good fit for low ranks of word frequency distributions (acknowledging some deviations at the very beginning of the ordering discussed in Tsonis, Schultz, and Tsonis (1997),Li (1998)) but little attention has been paid to the deviations in the tail. We will show that such deviations are much more important than expected.

2.2 Disagreements

One of the desirable properties of a law (as it happens with common physical laws) is to allow for accurate predictions.

The predicted number n of different words of a text formed by T words, can be obtained by applying Zipf's law and solving the following equation

$$\frac{1}{T} = p_1 n^{-\alpha} \tag{2.1}$$

where $1/T$ is the lowest probability that can be achieved by a word in a text of size T (Nowak, 2000a) and p_1 is the frequency of the most frequent word. From Eq. 2.1 we obtain

$$n = [Tp_1]^{1/\alpha} \approx Tp_1 \tag{2.2}$$

We processed [1] $T \approx 9 \cdot 10^7$ words of the British National Corpus (BNC) a corpus of modern English, both spoken (10%) and written (90%) (Appendix A).

[1] Words different than proper noun were lowercased. Marks were excluded. Inflected forms of the same (root) word were treated as different words.

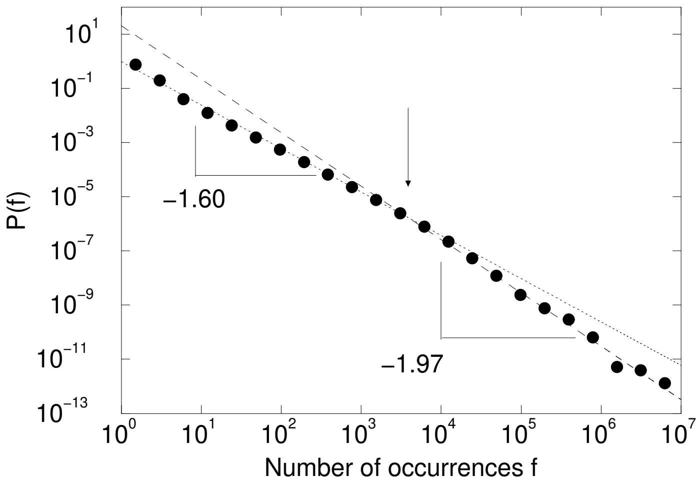

Figure 2.1: Probability that a word occurs f times. The first and the second power law decays have exponent $\beta_1 = 1.6 \pm 0.04$ and $\beta_2 = 1.97 \pm 0.06$, respectively. Statistics on the BNC ($T \sim 9 \cdot 10^7$ words, $n \sim 588,030$)

We obtained $P(1) = 0.0601046$, $\alpha = 1$ (power law regression). Unfortunately, $n = 588,030$ was very far from $\hat{n} \equiv 5.6 \cdot 10^6$. The big deviation observed could be attributed to a poor statistics or a bad fitting of the parameters intervening in the prediction, p_1 and α. We will show that there is a deeper reason.

We computed the probability density function of the frequency (in number of occurrences) of the BNC. More precisely, $P(f)$, the probability a word occurs f times in the corpus. The left half of the plot, shown in Figure 2.2, revealed a well-defined power law relationship between $P(k)$ and k whose exponent was $\beta = 1.5$. The value obtained was 1.6, but removing the two first points, corresponding to the most uncommon words, and thus corresponding to the frequencies being the most difficult to estimate, 1.5 was obtained ($\beta = 1.52 \pm 0.008$). In contrast, Eq. B.5 predicted $\beta = 2$. In addition, the plot of the probability density function in Figure 2.2 was specially clear. A question of bad statistics or fitting again?

2.3 Rethinking the law

A more careful sight of the rank ordering plot on our data revealed the existence of two different exponents in the same rank ordering plot (Figure 2.1). $\alpha_1 = \alpha \approx 1$ and $\alpha_2 \approx 2$ seem appropriate for ranks $i < i^* \in (10^3, 10^4)$ and $i \geq i^*$, respectively. Thus, the frequency of words becomes a double law, the initial Zipf's law and a steeper decay,

$$P(i) = \begin{cases} p_1 i^{-\alpha_1} & \text{if } i < i* \\ \sim i^{-\alpha_2} & \text{otherwise} \end{cases} \tag{2.3}$$

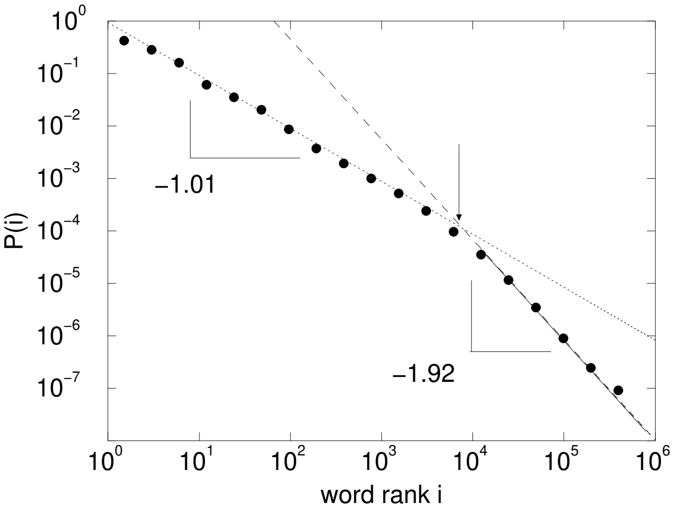

Figure 2.2: Probability of a word as a function of its rank i, $P(i)$. The first and the second power law decays have exponent $\alpha_1 = 1.01 \pm 0.02$ and $\alpha_2 = 1.92 \pm 0.07$, respectively. Statistics on the BNC ($T \sim 9 \cdot 10^7$ words, $n = 588,030$)

Let $x = [Tp_1(1)]^{1/\alpha_1}$. According to 2.3 and being $1/T$ the smallest probability, the number of different words predicted is

$$\hat{n} = \begin{cases} [Tp_1]^{1/\alpha_1} & \text{if } Tp_{i^*} < 1 \\ i^* \, [Tp_{i^*}]^{1/\alpha_2} & \text{otherwise} \end{cases} \tag{2.4}$$

where p_i is the frequency of the i-th most frequent word, $p_{1,000} = 1.06292 \cdot 10^{-4}$, $p_{5,000} = 1.71864 \cdot 10^{-5}$ and $p_{6,000} = 1.34702 \cdot 10^{-5}$.

The value of $\hat{n}$ calculated through Eq. 2.4 is $213,570$, much closer to the real value. Figure 2.3 shows the value of n, $\hat{n}$, obtained through Eq. 2.2) and 2.4; $i^* \approx 6,000$) and Ebeling/Pöschel approximation (Ebeling and Pöschel, 1994) as a function of T.

2.4 Discussion

A single slope $\alpha = 1$ can only be attributed to a superficial look on small-sized texts in which deviations in the tail of the distribution (of the rank-ordering plot) were attributed to finite size effects instead of a different exponent. Many previous work on English was performed on relatively small texts, i.e. $260,430$ words (Zipf, 1972a), $59,498$ words (Casti, 1995), $20,000$ words (Tsonis, Schultz, and Tsonis, 1997), far from the $\approx 9 \cdot 10^7$ words of the *BNC* we processed.

For long texts, the number of different words is mainly due to the second expression in Eq. 2.4. A relation $n \propto T^{-1/\alpha_2}$ was previously shown in (Ebeling

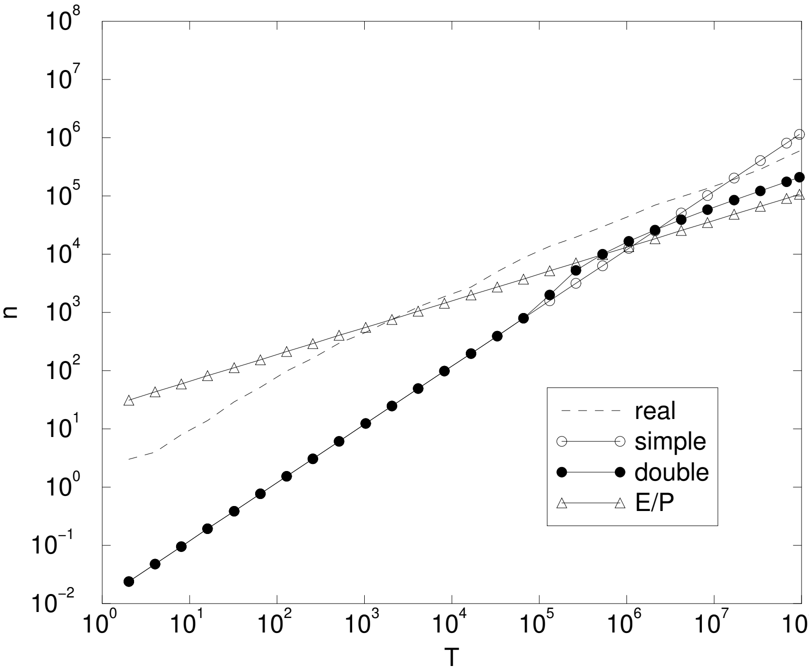

Figure 2.3: Number of different words, n, as a function of the total number of words in the sample, T. The real number is accompanied by estimations performed with the Zipf's law (Eq. 2.2), the two regime frequency observation (Eq. 2.4; $i^* \approx 6,000$) and the Ebeling/Pöschel approximation (Ebeling and Pöschel, 1994).

and Pöschel, 1994). More precisely, $n = 22.8T^{0.46}$.

The two observed exponents divide words in two different sets: a core lexicon formed by $\approx i^*$ versatile words and an unlimited lexicon for specific communication, the peripheral lexicon. We suggest that the size of the core lexicon is related with the average amount of words that human brain is able to efficiently store and use (Chapter 3 defines such efficient word storage and use) and also probably with minimum frequency allowing a word to spread in a communitity of speakers, as (Nowak, 2000a) shows. Words with the highest rank are very specyfic and obviously not shared by all speakers. According to the intersection of the lines aproximating the two regimes of $P(i)$ in Figure 2.1, the core lexicon of the BNC would be formed by 5,000-6,000 words. We do not mean that such size should be the same for whatever corpus or language. One must bear on mind that the estimation of the core lexicon size is made visually. Further work should be carried out in order to perform more accuarate estimations.

The existence of a core lexicon raises the question of how small can be a lexicon without drastically empoverishing communication. Pidgin languages provide examples of very small lexica. Estimates of the number of items of a pidgin vary from about $300 - 1500$ words, depending on the language (Romaine, 1992; Romaine, 1988). The number of lexical items of a speaker of an ordinary language is about $25,000-30,000$ (clearly not enough for the more than $500,000$ different words of the BNC) [2] while this amount is $1,500$ for a Tok Pisin speaker. It has been argued that these $1,500$ words can be combined into phrases so as to say anything that can be said in English (Hall, 1953). As expected, words of such small lexica are very multifunctional and a circumlocution is often recurred for covering the lexicon gaps. The transition from the exponent α_1 to α_2 takes place in the interval of rank $10^3 < i < 10^4$. We suggest that common languages also have a lexicon of the kind of pidgin languages, hidden by an unlimitited peripheral lexicon. Notice that although the size of the lexicon of a speaker can be very big, what counts for a successful communication are the words shared (stored and used) with the maximum number of speakers, that is, the words in the core lexicon.

The morphological simplicity and semantic generality that characterize pidgin and other known simplified lexica (Romaine, 1992) with regard to complex lexica can also be identified for the core lexicon. Table 2.1 summarizes them with examples from the BNC.

Some authors have pointed out the existence of two domains in the frequency of words (Naranan and Balasubrahmanyan, 1998), whose slopes agree with ours, or even three (Tuldava, 1996). Tuldava (1996) determined three slopes for the

[2] Although lexicon size estimates very often rely on roughly approximated counts, the Waring-Herdan's recursive model for frequency spectrum allows to perform more accurate counts. This model straighforwardly allows for the calculation of the number of words which are known by an author that do not appear in the sample, m_0. If L is the number of different word in the sample, it has been shown that A.H. Tammsaare's lexicon contained (by the time the sample was written) about $L + m_0 = 8,228 + 25,147 = 33,000$ words. See (Tuldava, 1996) for more details.

rank distribution in the following ranges:

$$\begin{aligned} i = 1 - 30 &\quad - \quad \alpha_1 = 0.7 \\ i = 30 - 1,500 &\quad - \quad \alpha_2 = 1.1 \\ i = 1,500 &\quad - \quad \alpha_3 = 1.4 \end{aligned}$$

Statistics were performed on A. H. Tammsaare's novel *Truth and Justice* and only lexemes were considered. The transition between the 2^{nd} and the 3^{rd} regime takes place in a rank closer to pidgin lexica size. Inflected forms of the same word were counted as different words in our statistics, suggesting that the rank at which the change in exponents takes place could be reduced. The slope of the less frequent words regime (1.4) is remarkably different than BNC's (2). Further study is needed for determining the origin of this disagreement.

We calculated the proportion of words of a text belonging to the core lexicon as a function of i^*, $S(i^*) = \sum_{i=1}^{i^*} P(i)$, being $P(i)$ the real probability of the i-th word) in order to illustrate the importance of the core. $S(1,000) = 0.69$, $S(4,000) = 0.84$, $S(5,000) = 0.86$ and $S(6,000) = 0.87$ show how recurring are such words. To sum up, the two frequency domains separate two clearly distinguishable word sets.

	core lexicon	peripheral lexicon
generality of terms	generic terms rather than specific terms (e.g. is_{9}, see_{96}, $group_{233}$, $live_{634}$, $know_{1,435}$ and $bird_{1,981}$)	larger vocabulary in a given domain (e.g. $biplane_{39,903}$, $coda_{43,482}$, $scarps_{68,727}$, $mycelium_{111,889}$, $anticoagulants_{113,286}$ and $microscopium_{432,607}$)
complexity of words	monomorphemic words (e.g. it_{7}, $made_{104}$, $year_{120}$, $hand_{246}$ and $mad_{3,312}$)	compounds (e.g. $airbrakes_{35,182}$, $fingerpriting_{53,988}$, $peachtree_{137,080}$, $breakdance_{163,284}$, $fingerlocks_{439,217}$ and $spillway_{453,615}$) and morphologically complex words (e.g. $childishly_{46,541}$, $literariness_{55,355}$, $thoughtlessness_{65,489}$, $overindebtedness_{97,885}$, $proletarianized_{103,707}$ and $multiculturated_{437,580}$)

Table 2.1: Comparison between the core lexicon and the peripheral lexicon. The intervening features were originally devised for comparing simple lexica (pidgin,creole,. . .) and complex lexica (Romaine, 1992). Example words are subindexed by its rank.

Chapter 3

Dual least effort

3.1 Introduction

When thinking how human language appeared on Earth, it seems reasonable to assume that our human ancestors started off with a communication system capable of rudimentary referential signaling, which subsequently evolved into a system with a massive lexicon, supported by a recursive system that could combine entries in the lexicon into an infinite variety of meaningful utterances (Hauser, 1996). In contrast, non-human repertoires of signals are generally small (Miller, 1981; Ujhelyi, 1996). We aim to provide new theoretical insights to the absence of intermediate stages between animal communication and language (Ujhelyi, 1996).

Here, we adopt the view that the design features of a communication system are the result of interaction between the constraints of the system and demands of the job required (Hauser, 1996). More precisely, we will understand the demands of such a task as providing easy-to-decode messages for the receiver. Our system will be constrained by the limitations of a sender trying to code such easy-to-decode message.

Many authors have pointed out that tradeoffs of utility concerning hearer and speaker needs appear at many levels. As for the phonological level, speakers want to minimize articulatory effort and hence encourage brevity and phonological reduction. Hearers want to minimize the effort of understanding and hence desire explicitness and clarity (Köhler, 1987; Pinker and Bloom, 1990). Regarding the lexical level (Zipf, 1972b; Köhler, 1987), the effort for the hearer has to do with determining what the word actually means. The higher the ambiguity (i.e. the number of meanings) of a word, the higher the effort for the hearer. Besides, the speaker will tend to choose the most frequent words. The availability of a word is positively correlated with its frequency. The phenomenon known as the *word frequency effect* (Gernsbacher, 1994) supports it. The most frequent words tend to be the most ambiguous ones. In fact, word ambiguity and word frequency are positively correlated (Reder, Anderson, and Bjork, 1974; Köhler, 1986).

Thereafter, the speaker tends to choose the most ambiguous words, which is opposed to the least effort for the hearer. G. K. Zipf referred to the lexical tradeoff as the *principle of least effort*. He pointed out it could explain the pattern of word frequencies but he did not give a rigorous proof of its validity (Zipf, 1972b). We saw in Chapter 1 that word frequencies obey the law called Zipf's law. Here we show that such a lexical compromise can be made explicit in a simple form of language game where minimization of speaker and hearer needs is introduced in an explicit fashion. As a consequence of this process, once a given threshold is reached, Zipf's law emerges spontaneously.

3.2 The model

In order to explicitly define the compromise between speaker and hearer needs, a cost function must be introduced. Given the nature of our systems, information theory provides the adequate mathematical framework (Ash, 1965). We consider a system involving a set of n signals $\mathcal{S} = \{s_1, ..., s_i, ..., s_n\}$ and a set of m objects of reference $\mathcal{R} = \{r_1, ..., r_i, ..., r_m\}$. The interactions between signals and objects of reference (hereafter objects) can be modeled with a binary matrix $\mathbf{A} = \{a_{ij}\}$, where $1 \leq i \leq n$ and $1 \leq j \leq m$. If $a_{ij} = 1$ then the i-th signal refers to the j-th object and $a_{ij} = 0$ otherwise. We define $p(s_i)$ and $p(r_j)$ as the probability of s_i and r_j, respectively. If synonymy was forbidden we would have

$$p(s_i) = \sum_j a_{ij} p(r_j) \tag{3.1}$$

since signals are used for referring to objects. We assume $p(r_i) = 1/m$ and $\omega_i \leq 1$ where $\omega_i = \sum_j a_{ji}$ is the number of synonyms of r_i in what follows. If synonymy is allowed, the frequency of an object has to be distributed among all its signals. The frequency of a signal, $p(s_i)$ is defined as

$$p(s_i) = \sum_j p(s_i, r_j) \tag{3.2}$$

According to the Bayes theorem we have

$$p(r_j, s_i) = p(r_j) p(s_i|r_j) \tag{3.3}$$

$p(s_i|r_j)$ is defined as

$$p(s_i|r_j) = a_{ij} \frac{1}{\omega_j} \tag{3.4}$$

Substituting Eq. 3.4 into Eq. 3.3 we get

$$p(r_j, s_i) = a_{ij} \frac{p(r_j)}{\omega_j} \tag{3.5}$$

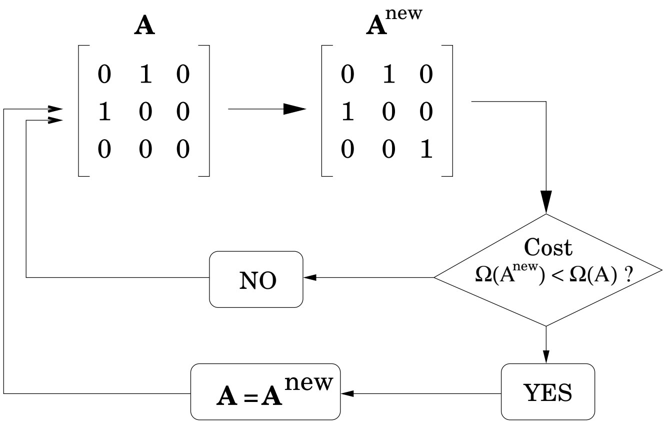

Figure 3.1: Basic scheme of the evolutionary algorithm used here. Starting from a given signal-object matrix $\mathbf{A}$ (here $n = m = 3$) the algorithm performs a change in a small number of bits (specifically, with probability ν, each a_{ij} can flip). The cost function Ω is then evaluated and the new matrix is accepted provided that a lower cost is achieved. Otherwise, we start again with the original matrix. At the beginning, $\mathbf{A}$ is set up with a fixed density ρ of ones.

The effort for the speaker will be defined in terms of the diversity of signals, here measured by means of the signal entropy, i. e.

$$H_n(\mathcal{S}) = -\sum_{i=1}^{n} p(s_i) \log_n p(s_i) \tag{3.6}$$

If a single word is used for whatever object, the effort is minimal and $H_n(\mathcal{S}) = 0$. When all signals have the smallest (non-zero) possible frequency, then the frequency effect is in the worst case for all signals. Consistently, $H_n(\mathcal{S}) = 1$.

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

$$H_m(\mathcal{R}|s_i) = -\sum_{j=1}^{m} p(r_j|s_i) \log_m p(r_j|s_i) \tag{3.7}$$

where $p(r_j|s_i) = p(r_j, s_i)/p(s_i)$ (by the Bayes theorem). The effort for the hearer is defined as the average noise for the hearer, that is

$$H_m(\mathcal{R}|\mathcal{S}) = \sum_{i=1}^{n} p(s_i) H_m(\mathcal{R}, s_i) \tag{3.8}$$

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

$$\Omega(\lambda) = \lambda H_m(\mathcal{R}|\mathcal{S}) + (1 - \lambda) H_n(\mathcal{S}) \tag{3.9}$$

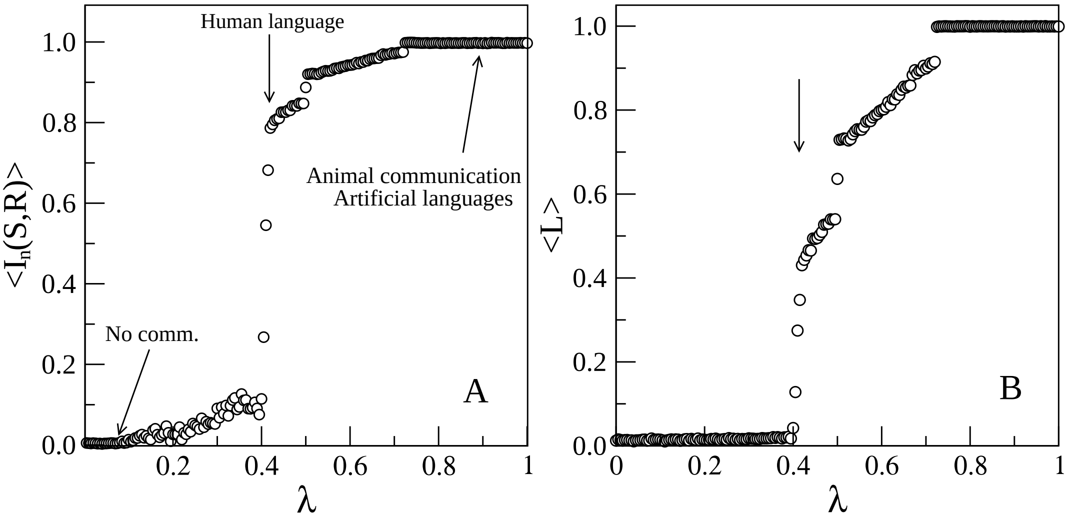

Figure 3.2: A. $< I_n(\mathcal{S}, \mathcal{R}) >$, the average information transfer as a function of λ. $\lambda^* = 0.41$ divides $< I_n(\mathcal{S}, \mathcal{R}) >$ into a no communication and perfect communication phase. B. Average (effective) lexicon size, $< L >$, as a function of λ. An abrupt change is seen for $\lambda \approx 0.41$ in both of them. Averages over 30 replicas, $n = m = 150$, $T = 2nm$ and $\nu = 2/\binom{n}{2}$.

where $0 \leq \lambda, H_n(\mathcal{S}), H_m(\mathcal{R}, \mathcal{S}) \leq 1$. The cost function depends on a single parameter λ which weights the contribution of each term.

3.3 Methods

$\Omega(\lambda)$ is minimized with the following algorithm, (summarized in Fig. 3.1). At each step, the graph is modified by randomly changing the state of some pairs of vertices and the new **A** - matrix is accepted if the cost is lowered (if an object has no signals, $\Omega(\lambda) = \infty$). The algorithm stops when the modifications on **A** are not accepted $T = 2nm$ times in a row. Configurations where an object has no signals assigned are forbidden.

3.4 Results

Two key quantities have been analyzed for different values of λ: the information transfer,

$$I_n(\mathcal{S}, \mathcal{R}) = H_n(\mathcal{S}) - H_n(\mathcal{S}|\mathcal{R}) \tag{3.10}$$

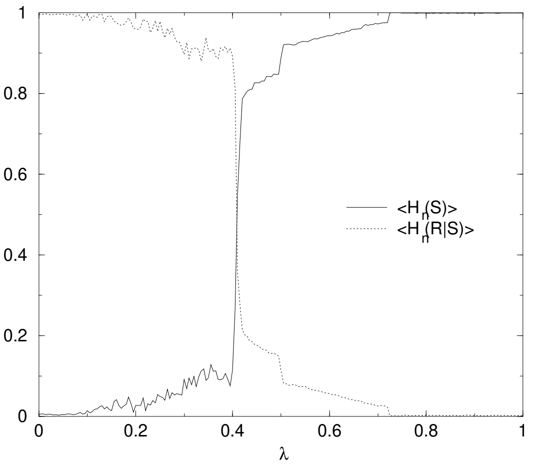

Figure 3.3: $\langle H_n(\mathcal{S})\rangle$ (solid line) and $\langle H_n(\mathcal{R}|\mathcal{S})\rangle$ (dotted line) versus λ (30 replicas). An abrupt change is found for $\lambda = \lambda^* = 0.41$.

which measures the accuracy of the communication, and the (effective) lexicon size, L, defined as

$$L = \frac{|\{i\,|\mu_i > 0\,\}|}{n} \tag{3.11}$$

where $\mu_i = \sum_j a_{ij}$ is the number of objects of s_i.

Three domains can be distinguished in the behavior of $I_n(\mathcal{S},\mathcal{R})$ versus λ, as shown in Fig. 3.2 A. First, $I_n(\mathcal{S},\mathcal{R})$ grows smoothly for $\lambda < \lambda^* \approx 0.41$. $I_n(\mathcal{S},\mathcal{R})$ explodes abruptly for $\lambda = \lambda^* \approx 0.41$. An abrupt change in L (Fig. 3.2 A) versus λ (Fig. 3.2 B) is also found for $\lambda = \lambda^*$. Single-signal systems ($L \approx 1/n$) dominate for $\lambda < \lambda^*$. Since every object has at least one signal, one signal stands for all the objects. $I_n(\mathcal{S},\mathcal{R})$ indicates that the system is unable to convey information in this domain. Rich vocabularies ($L \approx 1$) are found for $\lambda > \lambda^*$. Full vocabularies are attained beyond $\lambda \approx 0.72$. The maximal value of $I_n(\mathcal{S},\mathcal{R})$ indicates that the associations between signals and objects are one-to-one maps.

As for the energy function and directly related quantities, $H_n(\mathcal{S})$ is minimal for $\lambda < \lambda^*$ and becomes suddenly maximal for $\lambda > \lambda^*$. $H_n(\mathcal{R}|\mathcal{S})$ behaves inversely (Fig. 3.3). Thereafter, we have have $\Omega(\lambda) = \lambda$ for $\lambda < \lambda^*$ and $\Omega(\lambda) = -\lambda + 1$ for $\lambda > \lambda^*$, provided that λ is far enough from λ^* in both cases (Fig. 3.4).

As for the signal frequency distribution in every domain, very few signals have non-zero frequency for $\lambda < \lambda^*$ (Fig. 3.5 A), scaling consistent with Zipf's law appears for $\lambda = \lambda^*$ (Fig. 3.5 B) and an almost uniform distribution is obtained for $\lambda > \lambda^*$ (Fig. 3.5 C). As it occurs with other complex systems (Solé et al., 1996) the presence of a phase transition is associated with the emergence

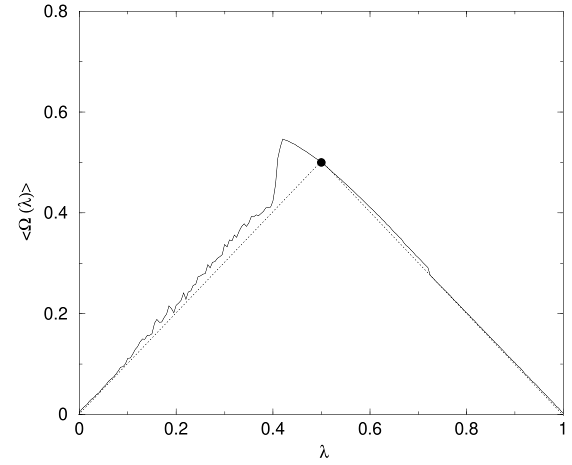

Figure 3.4: $\langle\Omega(\lambda)\rangle$, the mean energy function versus λ (30 replicas). $< \Omega(\lambda) >$ is a linear function of λ except for $\lambda = \lambda^* = 0.41$ and nearby values.

of power laws (Binney et al., 1992).

Knowing that $I_n(\mathcal{S}, \mathcal{R}) = I_n(\mathcal{R}, \mathcal{S})$ and using Eq. 3.10, minimizing Eq. 3.9 is equivalent to minimizing

$$\Omega(\lambda) = -\lambda I_n(\mathcal{S}, \mathcal{R}) + (1-\lambda)H_n(\mathcal{S}) \tag{3.12}$$

Other functions could be proposed. Interestingly, the symmetric version of Eq. 3.9 with conditional entropies in both terms of the right side

$$\Omega(\lambda) = \lambda H_m(\mathcal{R}|\mathcal{S}) + (1-\lambda)H_n(\mathcal{S}|\mathcal{R}) \tag{3.13}$$

will help us to understand the origins of the sharp transition. While the global minimum of $H_n(\mathcal{S})$ (one signal for all objects) is a maximum of $H_m(\mathcal{R}|\mathcal{S})$, the global minimum of $H_m(\mathcal{R}|\mathcal{S})$ (signal-object one-to-one maps with $n = m$) is a maximum of $H_n(\mathcal{S})$ in Eq. 3.9. Both terms of Eq. 3.9 are thus in conflict. In contrast, the global minimum of $H_n(\mathcal{S}|\mathcal{R})$ is a subset of the global minimum of $H_m(\mathcal{R}|\mathcal{S})$ in Eq. 3.13. Consistently, numerical optimization of Eq. 3.13 shows no evidence of scaling for Eq. 3.13. Not surprisingly, the minimization of Eq. 3.13 is equivalent to

$$\Omega(\lambda) = -I_n(\mathcal{S}, \mathcal{R}) + (1-\lambda)H_n(\mathcal{S}) \tag{3.14}$$

Notice that λ is present in only one of the terms of the right side of the previous equation. Zipf's hypothesis was based on a tension between unification and diversification forces (Zipf, 1972b) that Eq. 3.13 does not accomplish. Eq. 3.9 does.

3.5 Discussion

Theoretical models support the emergence of complex language as the result of overcoming error limits (Nowak and Krakauer, 1999) or thresholds in the amount of objects of reference that can be handled (Nowak, Plotkin, and Jansen, 2000). In spite of their power, these models make little use of some well known, quantitative regularities displayed by most human languages, such as Zipf's law (Zipf, 1972b; Miller and Chomsky, 1963). Most authors, however, make use of Zipf's law as a null hypothesis with no particular significance (Nowak, Plotkin, and Jansen, 2000). As far as we know, there is no compelling explanation for Zipf's law, although many have been proposed (Mandelbrot, 1966; Simon, 1955; Pietronero et al., 2001; Nicolis, 1991; Naranan and Balasubrahmanyan, 1998). See Chapter 5. Intermittent silence (random combinations of letters and blanks) reproduces Zipf's law (Miller, 1957; Li, 1992; Mandelbrot, 1966; Cohen, Mantegna, and Havlin, 1997) and are generally regarded as null hypothesis (Miller and Chomsky, 1963). Although intermittent silence and real texts differ in many aspects (Chapter 5), the possibility that Zipf's law results from a simple process (not necessarily intermittent silence (Miller and Chomsky, 1963)) has not been soundly denied. Our results show that Zipf's law is the outcome of the non-trivial arrangement of word-concept associations adopted for complying hearer and speaker needs. Sudden changes in Fig. 3.2 and the presence of scaling (Fig. 3.5 B) strongly suggest a phase transition is taking place at $\lambda = \lambda^*$ (Binney et al., 1992).

Maximal information transfer (that is, a one-to-one signal-object maps) beyond the transition is the general outcome of artificial life language models (Steels, 1996; Nowak, Plotkin, and Krakauer, 1999) and the case of animal communication (Deacon, 1997) where small repertoires of signals are found (Miller, 1981; Ujhelyi, 1996). The rather uniform shape of signal the frequency distribution for $\lambda > \lambda^*$ (Fig. 3.5 C) is consistent with the fact that the non-human species studied in 5 use their repertoires more uniformly than expected for Zipf's law. From the one hand, speaker constraints ($\lambda < \lambda^*$) are likely to cause species with a powerful articulatory system (providing them with a big potential vocabulary) to have a referentially useless communication system (Miller, 1981). From the other hand ($\lambda > \lambda^*$), least effort for the hearer forces a species to have a different signal for each object at the maximum effort for the speaker expense, which allows us to make the following predictions. First, non-human repertoires must be small in order to cope with maximum speaker costs. Consistently, their size is of the order of 20-30 signals for the larger repertoires (Miller, 1981). Second, the large lexicons used by humans can not be one-to-one maps because of the word frequency effect (Gernsbacher, 1994) that makes evident how lexical access-retrieval cost is at play in humans. Third, large lexicons with one-to-one maps can only be obtained under idealized conditions when effort for the speaker is neglected. This is the case of artificial language communication models, that reach maximal values of $I_n(\mathcal{S}, \mathcal{R})$ making use of fast memory access and the (theoretically) unlimited memory storage of computers (Steels, 1996; Nowak, Plotkin, and Krakauer, 1999).

$\lambda > \lambda^*$ implies not taking into account the speaker's effort. Getting the right word for a specific object may become unaffordable beyond a certain vocabulary size. Furthermore, a one-to-one map implies the number of signals has to grow accordingly as the number of objects to describe increases (when $m \to \infty$) and lead to a referential catastrophe. A referential catastrophe is supported by the statistics of human-computer interactions where the largest vocabularies follow Zipf's law (Ellis and Hitchcock, 1986) and are associated with a higher degree of expertise of the computer user. As the repertoire of potential signals is exhausted, strategies based on the combination of simple units are encouraged. Such a catastrophe could have motivated word formation from elementary syllables or phonemes but also syntax through word combinatorics. In a different context, some authors have shown that natural selection favors word formation or syntax when the number of required signals exceeds a threshold value (Nowak, Plotkin, and Jansen, 2000). We show that arranging signals according to Zipf's law is the optimal solution for maximizing the referential power under effort for the speaker constraints. Moreover, almost the best $I_n(\mathcal{S},\mathcal{R})$ is achieved before being forced to use one-to-one signal-object maps (Fig. 3.2). While other researchers have shown how overcoming phase transitions could have been the origin of the emergence of syntax (Nowak and Krakauer, 1999), our results suggest that early human communication could have benefited from remaining in a referential phase transition. There, communication is optimal with regard to the trade-off between speaker and hearer needs. An evolutionary prospect is that the number of objects to describe can grow keeping the size of the lexicon relatively small at the transition.

Having determined the only three optimal configurations resulting from tuning speaker and hearer requirements, the path towards human language can be hypothetically traced. First, a transition from a no communication phase ($\lambda < \lambda^*$) to a perfect communication phase providing some kind of rudimentary referential signaling ($\lambda > \lambda^*$). Second, a transition from a communication phase to the edge of the transition ($\lambda = \lambda^*$) where vocabularies can grow affordably (in terms of the speaker's effort) when $m \to \infty$. The latter step is motivated by (a) the positive correlation between brain size and cognitive skills in primates (where m can be seen a simple measure of them) (Reader and Laland, 2002). Humans may have had a pressure for economical signaling systems (given by large values of m) that other species did not have. The above mentioned emergence of Zipf's law in the usage of computer commands (the only evidence known of evolution towards Zipf's law, although the context are not human-human interactions) is associated with larger repertoires (Ellis and Hitchcock, 1986), suggesting that there is a minimum vocabulary size and also, because of the one-to-one mapping imposed by the speaker, a minimum number of objects encouraging Zipf's law arrangements.

Our results predict that no natural intermediate communication system can be found between small-sized lexica and rich lexica unless Zipf's law is used (3.2 B). This might explain why human language is unique with regard to other species, but not only so. One-to-one maps between signals and objects are the distinguishing feature of index reference (Deacon, 1997). Symbolic com-

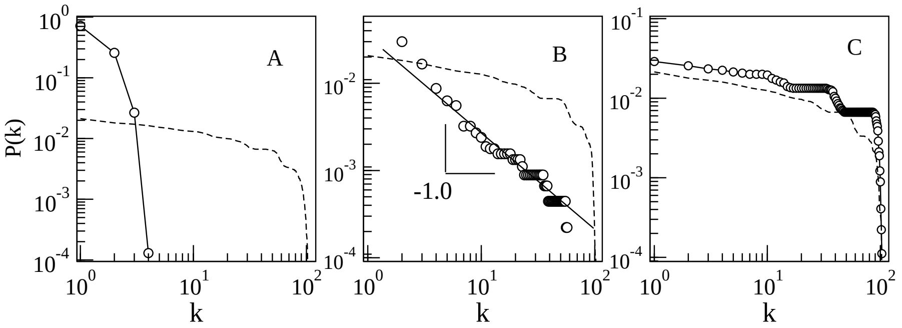

Figure 3.5: Signal normalized frequency, $P(k)$ versus rank, k, for (A) $\lambda = 0.3$, (B) $\lambda = \lambda^* = 0.41$ (B) and (C) $\lambda = 0.5$ (averages over 30 replicas, $n = m = 150$ and $T = 2nm$). Dotted lines show the distribution that would be obtained if signals and objects connected following a Poissonian distribution of degrees with the same number of connections of the minimum energy configurations. The distribution in (B) is consistent with human language ($\alpha = 1$).

munication is a higher-level reference in which reference results basically from interactions between signals (Deacon, 1997). Zipf's law appears on the edge of the indexical communication phase and implies polysemy. The latter is the necessary (but not sufficient) condition for symbolic reference (Deacon, 1997). Our results strongly suggest that Zipf's law is required by symbolic systems.

Chapter 4

Decoding least effort plus MaxEnt

4.1 Introduction

We assume a general communication framework where signals are mapped to the objects they refer to (Nowak and Krakauer, 1999; Nowak, Plotkin, and Krakauer, 1999) in this chapter. For vervet monkeys, we have alarms calls as signals and predators as objects (Seyfarth, Cheney, and Marler, 1980b). For human language, we have words as signals and meanings as objects, acknowledging that meaning is a complex matter to define (Ravin and Leacock, 2000b) and we humans make use of symbolic reference and not indexical reference as many animals seem to do (Hauser, 1996). For Unix computer commands, we have commands and their options as signals and the computer operations they imply as objects (Ellis and Hitchcock, 1986). For the immune system, we have reactivity patterns as signals and antigens as objects (Burgos, 1996; Burgos and Moreno-Tovar, 1996). We assume communication takes place between an ideal sender (speaker) and an ideal receiver (hearer). The task of the sender is to code an object using a signal that the receiver has to decode (Ash, 1965).

We typically have $\beta \approx 2$ for Zipf's law (Balasubrahmanyan and Naranan (1996); Chapter 2) but slight variations around β have been recorded (Balasubrahmanyan and Naranan, 1996). There are some interesting clear deviations:

1. Schizophrenia with $1 < \beta < 2$ (Zipf, 1972a).

2. Variations in the exponent when focusing on certain types of words (Fig. 4.1). We find $\beta = 3.35$ for English nouns (Fig. 4.1 B) [1] whereas we find $\beta = 1.94$ for English verbs (Fig. 4.1 A)

[1] Frequencies obtained from A. Kilgarriff's word-frequency list of the British National corpus (http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/bnc-readme.html).

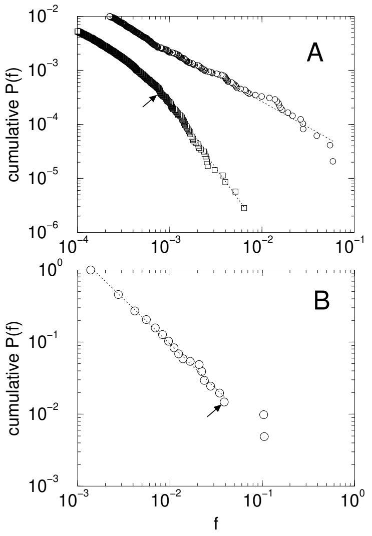

Figure 4.1: $P(f)$ the probability a signal has normalized frequency f in cumulative form. Power approximations are shown for every series (dotted lines). Arrows indicate the point considered as the end of the straight line for calculating the exponents β. A. p_f for English verbs with $\beta = 1.94 \pm 0.003$(circles) and English nouns with $\beta = 3.35 \pm 0.02$(squares). The core lexicon starts slightly before $f \approx 10^{-3}$ B. Unix computer commands issued by an experienced user $\beta = 2.24 \pm 0.028$

3. The peripheral lexicon (Chapter 2). Studies on multiauthor collections of texts show two domains in p_f. One domain with with $\beta \approx 2$ for the most frequent words and another domain with $\beta \approx 3/2$ for the less frequent words. The two regimes are said to divide words into a core and peripheral lexicon, respectively. We assume we focus on the core lexicon in the present chapter.

4. Shakespearean ouvres with $\beta = 1.6$ (Balasubrahmanyan and Naranan, 1996). This a rather controversial situation because Shakespearean works are likely to be a case of multiauthorship (Michell, 1999) and thus show the shape of a peripheral lexicon. What follows must be cautiously interpreted for this case.

Besides human language, scaling consistent with Zipf's law is found in the frequency of immune reactivity patterns (Burgos, 1996; Burgos and Moreno-Tovar, 1996) and the computer commands issued by experienced Unix users with $\beta = 2.24$ (Ellis and Hitchcock (1986); Fig. 4.1 B[2]).

[2] Statistics performed on the *bash* history file of an anonymous experienced user at the Complex Systems Lab.

We are aimed at answering the following questions:

1. Is there any general principle allowing to explain whatever form scaling in signal frequency distributions?

2. Is such a principle totally different from any explanation for the typical $\beta \approx 2$?

3. How does information transfer depends on β?

Many explanations have been proposed for scaling in word frequencies (Chapter 5). All the explanations for Zipf's law (except that in Chapter 3) forget a fundamental reason for which words are used: words are used according to their meaning. Real sentences are not a collection of words entirely chosen at random as many models intend (Simon, 1955; Mandelbrot, 1953; Miller, 1957; Li, 1992). Following the approach in Chapter 3, we assume that words are chosen according to their meaning and that the frequency of a word is a function of the objects eliciting it.

It has been shown that G. K. Zipf's proposal of a principle of least effort for the hearer and the speaker can explain $\beta \approx 2$ (Chapter 2). In a few words, G. K. Zipf proposed that Zipf's law results from a trade-off between hearer and speaker needs. In G. K. Zipf's rough intuition, the sender prefers a few words for all meanings (unification) and the hearer needs every meaning has a different word (diversification). The higher the degree of satisfaction of the needs of one of them, the less its effort. The model in Chapter 3 uses a parameter λ for minimizing $\Omega = \lambda E_D + (1-\lambda)E_C$, a linear combination of E_D, the coding effort (the effort for the hearer/receiver) and E_C, the coding effort (the effort for the speaker/sender), with $0 \leq \lambda \leq 1$. Sender and receiver needs are totally satisfied when $\lambda = 0$ and $\lambda = 1$, respectively. A phase transition separates a no communication phase (sender's full satisfaction) and a perfect communication phase (receiver's full satisfaction). Scaling consistent with Zipf's law with $\beta \approx 2$ is found at some intermediate $\lambda = \lambda^*$. We will refer to this model as the *dual least effort satisfaction model.* Here we show that scaling in word frequencies can be explained only complying with receiver needs under a convenient maximization principle. We will refer to this model as the *decoding least effort model.*

4.2 The model

We assume we have a set of signals $S = \{s_1, ..., s_i, ..., s_n\}$ and a set of objects of reference $R = \{r_1, .., r_j, ..., r_m\}$. We define a matrix of signal-object associations $A = \{a_{ij}\}$ $(1 \leq i \leq n\ ,\ 1 \leq j \leq m)$ where $a_{ij} = 1$ if the i-th signal and the j-th object are connected and $a_{ij} = 0$ otherwise. Here, we define the joint probability of the i-th signal and the j-th object as

$$p(s_i, r_j) = \frac{a_{ij}}{\sum_{k=1}^{n} \mu_k}$$

where μ_i, the number of objects linked to the i-th signal, is defined as

$$\mu_i = \sum_{j=1}^{m} a_{ij}. \tag{4.1}$$

Knowing the frequency of the i-th signal is

$$p(s_i) = \sum_{j=1}^{m} p(s_i, r_j) \tag{4.2}$$

we obtain

$$p(s_i) = \frac{\mu_i}{\sum_{k=1}^{n} \mu_k}.$$

The probability of understanding r_j when s_i is received is

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

so we have

$$p(r_j|s_i) = \frac{a_{ij}}{\mu_i}. \tag{4.3}$$

The probability definitions used here are simpler than in Chapter 3.

We define $\mathcal{H}$, the entropy of the number of objects per signal, as

$$\mathcal{H} = -\sum_{k=1}^{m} p_k \log p_k$$

where

$$p_k = \frac{|\{i|\mu_i = k \ and \ 1 \leq i \leq n\}|}{n}.$$

The maximization principle we will use for E_D comes from the observation that $\mathcal{H}$ is maximal at the point where scaling is found in Chapter 3 (Fig. 4.2). Thus, we can obtain $\{p_k\} = (p_1, ..., p_k, ..., p_m)$ using the maximum entropy principle (Kapur, 1989a; Montroll and Shlesinger, 1983). We define $\Phi(k)$ as the decoding effort (effort for the receiver) implied once a signal linked to k objects has been issued. We seek $\{p_k\}$ maximizing the *a priori* uncertainty $\mathcal{H}$ under the decoding effort we define here as

$$E_D = \frac{1}{n} \sum_{i=1}^{n} \Phi(\mu_i) \tag{4.4}$$

and the normalization constraint

$$\sum_{k=1}^{m} p_k = 1. \tag{4.5}$$

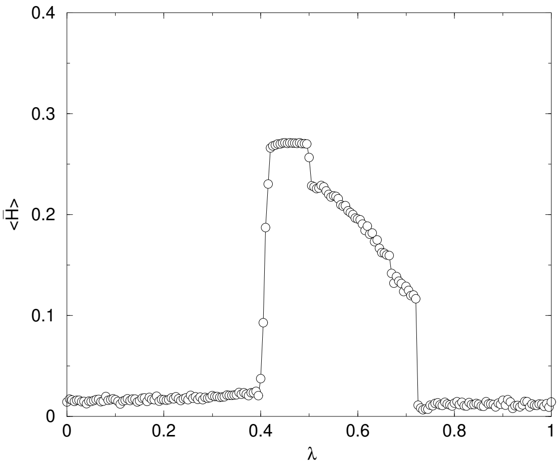

Figure 4.2: $< \bar{\mathcal{H}} >$ the mean normalized entropy of the number of objects per signal (solid line) versus λ, where $\bar{\mathcal{H}} =< \mathcal{H} > / \log m$. For $\lambda \approx 0.41$ as sharp transition takes place and scaling is found in the dual least effort model (n=m=150).

Rewriting Eq. 4.4 as

$$E_D = \sum_{k=1}^{m} p_k \Phi(k) \tag{4.6}$$

we end up with the functional

$$\Omega = \mathcal{H} + \alpha \sum_{k=1}^{m} p_k + \beta \sum_{k=1}^{m} p_k \Phi(k).$$

The distribution $\{p_k\}$ maximizing $\mathcal{H}$ will be deduced from the condition $\partial\Omega/\partial p_k = 0$ which leads to different distributions depending on Φ. Once s_i has been issued, the receiver must avoid interpreting an object that was not intended by the sender. The simplest way of satisfying the receiver needs is just minimizing μ_i, which leads to $\Phi(k) = k$ when $\mu_i = k$. A more sophisticated strategy consists of minimizing $H(R|s_i)$, the entropy of objects when s_i is given, defined as

$$H(R|s_i) = -\sum_{j=1}^{n} p(r_j|s_i) \log p(r_j|s_i) \tag{4.7}$$

$H(R|s_i)$ measures the uncertainty associated to the interpretation of s_i.

Replacing Eq. 4.3 into Eq. 4.7 we get

$$H(R|s_i) = -\sum_{j=1}^{m} \frac{a_{ij}}{\mu_i} \log \frac{a_{ij}}{\mu_i}$$

which gives $H(R|s_i) = \log \mu_i$. According to $H(R|s_i)$, if the i-th word has $\mu_i = k$ objects, then it implies an effort $\Phi(k) = \log k$.

For $\Phi(k) = k/m$ and large m, $\frac{\partial \Omega}{\partial p_k} = 0$ leads to (Haken, 1979)

$$p_k \sim e^{-k/<k>} \tag{4.8}$$

where c is a normalization term. For $\Phi(k) = \log k$, we obtain (Kapur, 1989a)

$$p_k \sim k^{\beta'}. \tag{4.9}$$

$\beta' < 0$ is satisfied provided that (Kapur, 1989a)

$$\frac{\sum_{k=1}^{m} k^{\beta'} \log k}{\sum_{k=1}^{m} k^{\beta'}} < \frac{1}{m} \sum_{k=1}^{m} \log k.$$

Zipf's law (Zipf, 1972a) can be straightforwardly obtained from Eq. 4.9 with $\beta' = -2$. If f is the frequency of a signal and p_f is the probability of f, Eq. 4.2 can be written as

$$f = \frac{k}{n \sum_{i=1}^{m} p_k k} = \frac{k}{n \langle k \rangle}.$$

If $P(k = K)$ is the probability the random variable k (the number of objects per signal) is K then using $p_f = P(k = fn \langle k \rangle)$ with Eq. 4.9 we get

$$p_f \sim f^{\beta'}.$$

Using the same argument on Eq. 4.8 we obtain

$$p_f \sim e^{-nf}.$$

4.3 Discussion

We have seen that explaining a wide range of exponents for the scaling in word frequencies is a relaxation of a more restrictive principle, minimizing both the coding and decoding effort. The dual least effort satisfaction model predicts that all signals will tend to have the same frequency if only receiver needs are satisfied. The decoding least effort presented here, with scaling in word frequencies, does not contradict the dual least effort model. The decoding least effort model assumes what is a side-effect close to the phase transition in the dual least effort model, i.e. maximizing $\mathcal{H}$.

We have seen that the decoding least effort model with $\Phi(k) = \log k$ predicts without specifying the value of β. The dual least effort model shows scaling consistent with Zipf's law for $\lambda \approx \lambda^*$ (Chapter 3). When $\lambda < \lambda^*$, word frequencies obey

$$P(i) \approx \begin{cases} 1 & if\ i = 1 \\ 0 & otherwise \end{cases} \tag{4.10}$$

where $P(i)$ is the frequency of the i-th most frequent word. Eq. 4.10 can be rewritten as $P(i) \sim i^{-\alpha}$ with $\alpha \to \infty$. When $\lambda > \lambda^*$, word frequencies obey $p(i) \sim 1/n$ which can be rewritten as $P(i) \sim i^{-\alpha}$ with $\alpha = 0$. Knowing (see for instance Naranan (1992),Naranan and Balasubrahmanyan (1992a); Chapter 2)

$$\beta = \frac{1}{\alpha} + 1, \tag{4.11}$$

we can argue that Eq. 1.1 is always present in the dual least effort model, not only for the transition but also at the two phases. The typical Zipf's law in human language is a particular case of scaling with non-extreme exponents, since $P(i) \sim i^{-\alpha}$ is only monotonically decreasing (and thus $P(i)$ can be defined as the frequency of the i-th most frequent word) only when $\alpha \in [0, \infty)$.

Now, we will find a simple relationship between E_D and β. c, the normalization term in Eq. 4.9, can be approximated solving

$$c \int_1^m k^{-\beta} dk = 1 \tag{4.12}$$

which leads to

$$c \approx \frac{1-\beta}{m^{1-\beta} - 1} \tag{4.13}$$

provided $\beta \neq 1$. β can be approximately determined substituting Eq. 4.9 into the definition of E_D of Eq. 4.6 as follows

$$E_D = \int_1^m c k^{-\beta} \log k dk \tag{4.14}$$

Solving the integral in the right side of the previous equation with $\beta \neq 1$ we get

$$E_D = c\frac{1}{1-\beta}\left[m^{1-\beta}\left(\log m - \frac{1}{1-\beta}\right) + \frac{1}{1-\beta}\right] \tag{4.15}$$

which we rewrite as

$$E_D = \frac{1}{m^{1-\beta}-1}\left[m^{1-\beta}\left(\log m - \frac{1}{1-\beta}\right) + \frac{1}{1-\beta}\right] \tag{4.16}$$

using Eq. 4.13. Notice that the previous equation is undetermined for $\beta = 1$ or $m = 1$. If $m \to \infty$ and $\beta > 1$ we have

$$\beta = \frac{1}{E_D} + 1 \tag{4.17}$$

It follows from Eq. 4.11 and Eq. 4.17 that $E_D = \alpha$. Since $E_D \geq 0$ (when $\beta > 1$), then solving

$$\frac{dE_D}{d\beta} = -\frac{2}{(\beta-1)^2} = 0$$

gives a global minimum of E_D for $\beta \to \infty$.

Knowing that $\alpha = 0$ and therefore $\beta \to \infty$ minimize not only E_D but also maximize the potential information transfer (Cover and Thomas, 1991; Suzuki, Tyack, and Buck, 2003), we may ask why human language has chosen $\alpha \approx 1$ and therefore $\beta \approx 2$ as its typical exponents. Is the answer that human language is more a system of thought and mental representation than a communication system as some researchers have proposed (Chomsky, 1965a; Bickerton, 1990; Jackendoff, 1994)? Probably the answer is that the pressure for maximizing information transfer, minimizing the decoding effort in human language has to satisfy conflicting goals. The dual least effort model tells us that adding coding least effort is a suitable answer for $\beta \approx 2$. Other exponents require putting into consideration other constraints. Nouns, with $\beta \approx 3.35$ are closer to the theoretical maximum information limit ($\beta \to \infty$), suggesting they have violated the balance or maximum tension between coding and decoding needs in other to achieve higher information transfer, that is, lower decoding effort. Eq. 4.17 suggests that nouns have lower decoding effort than the typical E_D given by Zipf's law with $\beta \approx 2$. Similarly, schizophrenic speech with $1 < \beta \leq 2$ suggest they are not taking into account the effort for the hearer their exponents imply high values of E_D. This is consistent with the suspect that schizophrenic speakers tend to lump together too many meanings in one form of expression. Schizophrenics overload word meanings (Zipf, 1972a). Therefore, exponents are indicators of E_D and have to do with information transfer. To make it more explicit, Eq. 1.1 and Eq. 4.17 give

$$p_f \sim f^{\frac{1}{E_D}+1}.$$

It should be understood from the present work that $\beta < \infty$ does not imply that scaling in word frequency has nothing to do with effective communication although different mechanisms can lead to Zipf' law (Suzuki, Tyack, and Buck, 2003). Eq. 4.17 bridges the gap between power word frequency distributions and communicative efficiency. There are many possible ways of minimizing the decoding effort, but probably only one where hearer and speaker needs are at the maximum tension, i.e. $\beta \approx 2$.

Our work puts a step forward to understand complex and simpler communications systems. The former making use of $\Phi(k) = \log k$ and the latter $\Phi(k) = k$. Scaling in different contexts (Zipf, 1972a; Burgos, 1996; Burgos and Moreno-Tovar, 1996; Ellis and Hitchcock, 1986) suggests that many systems in nature make use of non-trivial mechanisms for reducing the uncertainty associated to the codes they generate. Minimizing $\Phi(k) = k$ helps to decrease the uncertainty associated to the interpretation of a signal but does not lead scaling.

Chapter 5

Zipf's law meaningfulness

5.1 Introduction

Many explanations have been proposed for Eq. 1.1 with $\beta \approx 2$. Given the amount of explanations and the simplicity of some explanations, answering to the following questions is necessary:

- Is Zipf's law meaningful *in human language*? The question requires some reflections. We could have said 'Is Zipf's law meaningful?' The answer to the latter question is obviously 'No' because too many different mechanisms can lead to Zipf's law. Therefore, the tail *in human language* is fundamental for the relevance of the question. Besides, we have not clarified what meaningfulness is. We will understand meaningfulness here as a deep relationship between Zipf's law and human language, as a communication system (weak meaningfulness, since other non-human species seem to have simple communication systems (Hauser, 1996), but probably not as complex as that of humans) or as language in the strict sense, that is, syntax (with recursion) (Hauser, Chomsky, and Fitch, 2002) or symbolic reference (Deacon, 1997). Alternative ways of defining meaningfulness are *some powerful and universal psychological force that shapes all human communication in a single mold* or a *intelligent* or *purposeful source* (Miller and Chomsky, 1963). We will discard *intelligent source* since intelligence is dissociated from language in the so-called *idiot savants* (Deacon, 1997). We will also discard *purpose* because it is too vague in the context of human language. Nonetheless, such *a powerful and universal psychological force* is in the spirit of the model Chapter 3. The soundness of the latter definition of meaningfulness will be supported by the present chapter.

- What are the requirements of an explanation for Zipf's law in humans? We hereafter say an explanation is valid if it satisfies such a set of requirements.

- How can the suitability of valid explanations be evaluated?

5.2 Hypothesis testing

Answering to the question 'Is Zipf's law meaningful in human language?' requires introducing some basic ideas of hypothesis testing (what follows is borrowed from (Easton and McColl, web page)). Answering to a certain question of interest is simplified into two competing claims or hypotheses between which we have a choice, i.e. the null hypothesis, denoted $\mathcal{H}_0$, against the alternative hypothesis, denoted $\mathcal{H}_1$. These two competing claims or hypotheses are not however treated on an equal basis: special consideration is given to the null hypothesis. We have two common situations:

- The experiment has been carried out in an attempt to disprove or reject a particular hypothesis, the null hypothesis, thus we give that one priority so it cannot be rejected unless the evidence against it is sufficiently strong. For example, $\mathcal{H}_0$: *'Zipf's law is not meaningful in human language'* against $\mathcal{H}_1$: *'Zipf's law is meaningful in human language'.*
- If one of the two hypotheses is *simpler* we give it priority so that a more *complicated* theory is not adopted unless there is sufficient evidence against the simpler one. For example, it is 'simpler' to claim that *Zipf's law is meaningful in human language* than it is to say that *it is meaningful.*

Hypotheses are often statements about population parameters like expected value and variance (Sokal and Rohlf, 1995). A hypothesis might also be a statement about the distributional form of a characteristic of interest. The outcome of a hypothesis test is *'Reject $\mathcal{H}_0$ in favour of $\mathcal{H}_1$'* or *'Do not reject $\mathcal{H}_0$*. It has been said that Zipf's law is a sort of null hypothesis (Miller and Chomsky, 1963; Nowak, 2000a). If $\mathcal{H}_0$ is Zipf's law, what is the hypothesis $\mathcal{H}_1$? Human language typically follows Zipf's law, as well as all the models reproducing Zipf's law. Where are the alternative metrics or distributions? Saying Zipf's law is a sort of null hypothesis in the context of mechanisms reproducing Zipf's law is inaccurate and therefore misleading. The only way of distinguishing alternative hypothesis in models reproducing Zipf's law is adding extra information.

We will use information about the mechanism leading to Zipf's law for discriminating between Zipf's law meaningfulness and Zipf's law absence of significance. We thus define a test T formed by two alternative hypothesis: $\mathcal{H}_1$, *'Zipf's law is meaningful in human language'* and a null hypothesis $\mathcal{H}_0$, *'Zipf's law is not meaningful in humans'.* A different test T' is erroneously defined in the literature (Wolfram, 2002; Li, 1992; Miller and Chomsky, 1963). T' is formed by $\mathcal{H}'_1 = \mathcal{H}_1$ and $\mathcal{H}'_0$, defined as *'Zipf's law is not meaningful'.* Notice the tail $\mathcal{H}'_0$ the tail *in human language* is omitted on purpose. Such asymmetric definition of the test is widespread in the literature.

The null hypothesis, $\mathcal{H}_0$, represents a hypothesis that has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved. Special consideration is given to the null hypothesis. This is due to the fact that the null hypothesis relates to the statement being tested, whereas the alternative hypothesis relates to the statement to be accepted if the null is rejected.

The final conclusion once the test has been carried out is always given in terms of the null hypothesis. We either *Reject* $\mathcal{H}_0$ *in favour of* $\mathcal{H}_1$ or *Do not reject* $\mathcal{H}_0$. We never conclude *Reject* $\mathcal{H}_1$, or even *Accept* $\mathcal{H}_1$. If we conclude *Do not reject* $\mathcal{H}_0$, this does not necessarily mean that the null hypothesis is true, it only suggests that there is not sufficient evidence against $\mathcal{H}_0$ in favour of $\mathcal{H}_1$. Rejecting the null hypothesis then, suggests that the alternative hypothesis may be true. The test is not only incorrectly formulated as T' instead of T, but also the interpretation of the outcome of T'. When $\mathcal{H}'_0$ is proven, then $\mathcal{H}_1$ is strongly suggested to be false:

- *"I suspect that in fact the law (Zipf's law in human language) has a rather simple probabilistic origin. Consider generating a long piece of text by picking at random from k letters and a space."* (from (Wolfram, 2002)).
- *"It is clear now that the existence of the Zipf's-law-like word frequency distribution in random texts (intermittent silence models here) is purely due to the choice of the rank as the independent variable (instead of rank)..." ... "This strongly suggests that the power law as expressed in natural languages is also purely due to the choice of the rank as the independent variable"* (from (Li, 1992)).

The previous erroneous suspects are both based on intermittent silence, a model that will be analyzed in depth in Section 5.3. In practice, when evolution of language models assume Zipf's law without a linguistic interpretation (Nowak, Plotkin, and Jansen, 2000; Nowak, 2000a; Nowak, 2000b), they assume $\mathcal{H}_1$ is false.

Sometimes, $\mathcal{H}_1$ is assumed to be implicitly true whereas $\mathcal{H}_0$ has not been proven or disproven. This is the case of (McCowan, Doyle, and Hanser, 2002) when comparing human words with dolphin whistles and squirrel monkey calls. Additionally, the work in (McCowan, Doyle, and Hanser, 2002) raises other methodological questions, such as the convenient technique that must be used for categorizing dolphin whistles (Janik, 1999).

Section 5.3 is a critical review of three selected classical models for Zipf's law and shows the difference between $\mathcal{H}_0$ and $\mathcal{H}'_0$. Section 5.4 summarizes existent Zipf's law models. Section 5.5 shows how $\mathcal{H}_0$ should be formulated and proposes a context suitable for human language where $\mathcal{H}_0$ is rejected. A discussion of the suitability of all models is given in Section 5.6. Using T and not T' we will conclude Zipf's law is meaningful.

5.3 Selected classic models

This section contains an in depth analysis of three classic models. The suitability of every model as a strict model of Zipf's law for human language or as a support for $\mathcal{H}_0$ is studied. It is important to bear on mind the aim of the present section is not enumerating all the differences between such models and real human language, since whatever model, even the best one, is a simplification of reality. Instead, emphasis is made on the essential differences between actual word use and the way words are used in such models. For instance, as pointed in (Balasubrahmanyan and Naranan, 1996), intermittent silence models imply an exponential distribution of word lengths against the more accepted Poisson-like (with a typical length greater than one) for real word lengths (Wimmer et al., 1994; Wimmer and Altmann, 1996). Even if intermittent silence models where adapted to solve such inconsistency, essential differences with regard to the underlying mechanisms behind real word use would still remain. The next subsections are devoted to understand such fundamental differences.

5.3.1 Intermittent silence

One of the most simple models reproducing Zipf's law is the following. Take a finite set of symbols $\Sigma = L \cup \{ \_\}$ where $\_$ stands for the blank space and L is the set of letters $L = \{a, b, ..., z\}$. Form a sequence by choosing elements L with probability $p/|L|$ and a blank with probability $1 - p$. Every time $\_$ is chosen, it marks the beginning of a new word and the end of the previous word (if it is the case).

Such a random process is called a *intermittent silence* (Miller, 1957), *monkey language* (Miller, 1957; Casti, 1995) or or simply a *random text* (Cohen, Mantegna, and Havlin, 1997; Li, 1992) model. The term *monkey language* latter name follows from the fact the random sequence would be the expected outcome of a monkey hitting a keyboard having Σ as keys). Hereafter we will use the term intermittent silence, since monkey language deceivingly invokes the communication of primates in the wild and random text is too general since other ways of generating a random text reproducing Zipf's law do exist (e.g. Simon's model (Simon, 1955)). The term random is often misused (Suzuki, Tyack, and Buck, 2003), or in our opinion, misinterpreted. Despite of the surprising suitability of intermittent silence, they are generally regarded as null hypothesis that whatever explanation has to face (Miller and Chomsky, 1963).

The process is not appropriate for proving $\mathcal{H}_0$ for two reasons:

- The source of words. We humans chose words from a finite mental lexicon (Carroll, 1994), in other words, we choose words from a finite set of words.

- The way words are chosen. We humans choose words according to their meaning.

If single-author texts, the vocabulary size of a sample text is bounded by a certain finite value. Such a value is the author's vocabulary size. A single speaker has a finite vocabulary. The most optimistic estimates give about 10^4 words for adult speakers (Miller and Gildea, 1987). Even in the largest counts, one has to keep on mind that a power behavior is not found for the less frequent words of a single author (Montemurro, 2001). What does it happen in multiauthor texts? The domain in which Zipf's law (that is, $\beta \approx 2$) holds in multiauthor texts is formed by about the 6000 most frequent words, the so-called core lexicon (Chapter 2).

Individual speaker have memory limitations and communities of speakers seem to have greater difficulties in obeying Zipf's law with $\beta \approx 2$.

Different mechanisms make the size of the lexicon theoretically infinite

- If we by a possible word mean a sequence of letters (or phonemes) without blanks, the number of words that can be formed is obviously infinite, even when taking into account that some combinations of letters (or phonemes) are not allowed as word because of orthographic (or phonetic) constraints.
- Languages have word derivation and compounding (Akmajian, 1995) for creating new words from existing ones.

It is important to understand that the lexicon is theoretically infinite but finite in practice, since

- Words have a characteristic length and fast decay in the length probability distribution for long lengths (Wimmer et al., 1994; Wimmer and Altmann, 1996; Riedemann, 1996).
- One thing is speaking using reusing derived or compound words from a mental lexicon and another thing is speaking creating new derived or compound words. Reusing is generally by far much more frequent than creation. If creation was larger, one should try to find if Zipf's law holds for creations before rejecting the previous statement.

Therefore, we will assume the mental lexicon is finite.

Speaking from a mental lexicon requires:

- The basic units must be words (and not letters). Intermittent silence models assume every word is created from scratch by combining letters. We humans choose from pre-existing words in the mental lexicon. We do not build them on the fly (besides inflectional variations and rather uncommon word derivations and compoundings). A model consistent with human language must work at the lexical level and not at lower levels ones. Therefore, only words are valid units. We will label this requirement as *Word*.
- Words must be chosen according to their meaning. Words have no meaning in intermittent silence. We will label this requirement as *Ref.*
- The lexicon must be finite. We will label this requirement as *Fin.* The lexicon size, that is, the number of different words, increases as the sequence grows in intermittent silence models.

Because of the neglecting the mental lexicon, intermittent silence and real texts manifest other differences such as,

- Vocabulary growth with regard to texts length is faster in intermittent silence than in real texts (Cohen, Mantegna, and Havlin, 1997).

- Intermittent silence fills the frequency spectrum using brute force and the way real texts fill it effortlessly (even when the probability of every letter is not the same but estimated from real texts) (Appendix C). Real texts fill the spectrum easily because the their process is confined to a bounded size lexicon.

- Zipf's law (with $\beta \approx 2$) in real word frequencies is independent of word length. Words of the same length follow Zipf's law (with $\beta \approx 2$) in human language. In contrast, intermittent silence (even using biased letter probabilities) fail (Appendix C).

5.3.2 Birth process

H. A. Simon proposed the following process for explaining Zipf's law (Simon, 1955). In each iteration step, the text grows by one word. The $(t+1)$ word will be either a new one (with probability ψ) or an old word (with probability $1-\psi$) that has already appeared in the text. The old word is obtained choosing one member of the sequence at random, that is, all occurrences of words in the sequence have the same probability of being chosen. The distribution of the process follows Eq. 1.1 with (Simon, 1955; Zanette and Manrubia, 2001)

$$\beta = 1 + \frac{1}{1-\psi} \tag{5.1}$$

See Appendix D for a simple proof.

Simon's model reproduces Zipf's law with $\beta \approx 2$ for small values of ψ. *Fin.* implies that that Simon's should be able of reproducing Zipf's law when $\psi = 0$, as predicted by Eq. 5.1 but this is not the case. For $\psi = 0$ the Simon process becomes a Polya process. For understanding what a Polya process is, *'think of an urn of infinite capacity to which are added balls of two possible colors-read and white, say. Starting with one red and one white ball in the urn, add a ball each time, indifenitely, according to the rule: choose a ball in the urn at random and replace it; if it is red, add a red; if it is white, add a white. Obviously, this process has increments that are path-dependent-at any time the probability that the next ball added is red exactly, equals the proportion of red.'...Polya proved in 1931 (Polya, 1931) that in a scheme like this the proportion of red balls does tend to a limit X, and with probability one. But X is a random variable uniformly distributed between 0 and 1.* (Arthur, 1994, p. 36). The result can be generalized to balls with more than two different colors. Then, equating different colors with different words, it can be shown that whatever $P(i)$ satisfying $\sum_i P(i) = 1$ is a solution of the such a generalized Polya process (Arthur, 1994), where n is the number of words. Therefore, $\alpha = 0$, predicts a uniform frequency versus rank distributions, which are a subset of the Polya process quasistationary solutions. Eq. 1.1 is not warranted for $\alpha = 0$. Therefore, Simon's model fails for *Ref.* and *Fin.* whereas it satisfies *Word.* Simon's model is based on introducing new words and the most of the words that are created or introduced are nouns. Unfortunately, such a process can not realistically explain the exponent of nouns, which is $\beta = 3.35$ (Ferrer i Cancho, 2003), since Eq. 5.1 leads to $\psi = 0.57$. It is hard to believe that we humans use a word from the pool only about 40% of the times. The Simon model, at least in its basic version, is not linguistically sound. Finite vocabularies are favoured.

5.3.3 Word length minimization

Mandelbrot (Mandelbrot, 1953) argued that communication using a repertoire of n words must maximize

$$H = -\sum_{i=1}^{n} p_i \log p_i \tag{5.2}$$

subject to a certain constraint

$$C = \sum_{i=0}^{n} C_i p_i \tag{5.3}$$

where p_i is the frequency of the i-th most frequent word and C_i is the cost associated to the i-th most frequent word. We define L_i as the the length of the word whose rank is i. All the possible words that can be formed concatenating letters are ranked according the number of letters they contain, $i < j$ if $L_i < L_j$, and an arbitrary rank is assigned when $L_i = L_j$. Such arbitrary rank must preserve ranks are a partial order with respect to L_i. It follows that $C_i \approx log_N i$, where N is the size of the alphabet. Using the Lagrange multipliers method, Eq. 1.2 is obtained. Mandelbrot's model fails for *Ref.* whereas satisfies *Word* and *Fin.*

Although H is the average information per signal, it has very little to do with effective communication. Rapoport (Rapoport, 1982) tried to misleadingly justify the link between Mandelbrot's model and information transfer. The average information transfer and the average information per signal are different measures. It is true, as Rapoport points out, that minimizing H leads to no communication, since $H = 0$ is reached when a word has probability 1. Nonetheless, the converse is not true, maximizing H does not lead to effective communication.

Shannon (1948) defined a measure of effective communication, i.e. information transfer. Assuming we have a set of signals (e.g. words) $S = \{s_1, ..., s_i, ..., s_n\}$ and a set of objects of reference $R = \{r_1, ..., r_j, ..., r_m\}$, the information transfer can be defined as

$$I(S, R) = H(S) - H(S|R) \tag{5.4}$$

where $H(S)$ is the entropy associated to signals, $H(R|S)$ is the entropy associated to objects when signals are known. $H(S)$ in Eq. 5.4 is the H that Rapoport (Rapoport, 1982) argues it is maximized in Mandelbrot's derivation. It follows from Eq. 5.4 that $H(S) \leq I(S, R)$. In other words, maximizing $H(S)$ only maximizes $I(R, S)$ when $H(S|R)$ is constant, a very particular case that it is not at all assumed in Mandelbrot's derivation.

Let us show an example where $H(S)$ is maximal but $I(S,R)$ is minimum. Imagine we have a set of signals S (e.g. words) and a set of objects of reference R. Imagine every signal in S is linked to every object of reference in R. Assume that the frequency of a signal is proportional to the number of objects of reference such signal is linked to. Then, it can be easily seen that $H(S)$ is maximal, but $I(S,R) = 0$ because $H(S|R)$ is also maximal ($H(S|R) = H(S)$). The latter is true because coding an object implies making a decision about which among the $n = |S|$ possible signals is the suitable (see Section 5.5 (Ferrer i Cancho, 2003) for the precise probability definitions required for calculating $I(R,S)$). To sum up, it is true that H maximizes the potential information transfer (Suzuki, Tyack, and Buck, 2003; Cover and Thomas, 1991), but just that. It has nothing to say about the actual information transfer.

Another problem of maximizing H is that if the language follows a Zipf's law and the lexicon is infinite then the entropy is not possible to determine from empirical data because the entropy function has its points of discontinuity exactly on the Zipf's laws with $\beta = 2$ (Harremoës and Topsøe, 2001; Harremoës and Topsøe, 2002). If $beta < 2$ one can make estimates of the entropy, but the closer β is to 2 the more difficult the estimation becomes (Antos and Kontoyiannis, 2001)[1].

5.4 A general summary

It would be difficult to describe briefly the remaining models that have been proposed for Zipf's law. Instead, existent models will be classified in order to capture the essential features needed by the present chapter. All existing models can be classified into tree major groups (Table 5.1):

A. Models based on some sort of optimization principle.

B. Models based on stability criteria. Here stability has not the meaning in the analysis of dynamical systems but a different one. (Harremoës and Topsøe, 2001; Harremoës and Topsøe, 2002) define a condition of stability where the H may be small or big for approximately equal probability distributions. Only Zipf's law (with $\beta \approx 2$) satisfies such condition. The similarity between distributions is calculated using the Kullback-Leibler divergence or relative entropy [2].

[1] Peter Harremoës, personal communication

[2] To measure how much an observed distribution P differ from a theoretical distribution Q I use the information divergence from P to Q denoted D(P,Q). This quantity is also called Kullback-Leibler divergence or relative entropy. Information divergence is not symmetric in its arguments so that D(P,Q) and D(Q,P) are different quantities. It is important that the observed distribution appears as first argument and that the theory appears as second argument. Only this allows us to have a finite lexicon for the observed distribution and an infinite lexicon for the theoretical distribution. The asymmetry reflects that one can only observe something which has positive probability, but one can have something which is not observed but has positive probability (Peter Harremoës, personal communication).

C. Models not based on any of the previous. Type C models are challenging since they they require the weakest assumptions. Type C models are of two kinds:

   C.1. Plain stochastic processes (e.g. intermittent silence (Miller, 1957; Mandelbrot, 1953; Mandelbrot, 1966; Li, 1992) and Simon's model (Simon, 1955)). Unfortunately, we have seen in Section 5.3 that none of them qualify for an explanation.

   C.2. Differential equations for $dP(i)/di$ or $dP(f)/df$ (Tuldava, 1996; Montemurro, 2001; Tsonis, Schultz, and Tsonis, 1997)

As for the type C.2 models, they are not explanations in the strict sense but tautologies, since they hypothesize how f or $P(f)$ vary but they do not clearly explain why the variation is the one chosen in a sound way. Once the variation is specified, the hypothetical distribution follows in a straightforward way. That is why we call them tautologies.

Type A models generally make use of the maximum entropy principle (Kapur, 1989b; Cover and Thomas, 1991) for obtaining the word frequency distribution. Different kinds of entropy measures are used: Shannon entropy (Chapter 3-4,Nicolis (1991), Naranan and Balasubrahmanyan (1992a) and Naranan and Balasubrahmanyan (1992b)), degenerate entropy (Naranan and Balasubrahmanyan, 1993), algorithmic entropy (Balasubrahmanyan and Naranan, 1996), Tsallis entropy (Denisov, 1997), and Rényi entropy (Bashkirov and Vityazev, 2000; Bashkirov, 2003). Some of the models examined here come from a very general framework (Denisov, 1997; Bashkirov and Vityazev, 2000; Bashkirov, 2003). Although we have pointed out in Section 5.5 that the linguistics context is very important for Zipf's law meaningfulness, such models are reviewed here since they suggest that Zipf's law in linguistics could be a consequence of very general principles.

Some models determine the exponent and some other do not. Models not determining the exponent are interesting because real values of β exhibit some degree of variation around 2 (Balasubrahmanyan and Naranan, 1996) with some rather exceptional interesting values in a linguistic context, such as $1 < \beta < 2$ for schizophrenia (Zipf, 1972a) and $\beta = 3.35$ for English nouns (Chapter 4). When the exponent is not determined, it can be determined using real data. An undetermined exponent can be regarded as a source of imprecision against the typical $\beta \approx 2$. The requirement that models determine the exponent will be labeled as *Deter*. It is important to notice that some models do not determine $\beta \approx$ but provide a narrow interval including $\beta \approx 2$. This is the case of (Bashkirov and Vityazev, 2000; Bashkirov, 2003). Some explanations may not allow determining the exponent using real data or further constraints. Good explanations must be testable (Popper, 1968; Medawar, 1969).

Since there are many explanations for Zipf's law, explanations allowing determining the exponent adding linguistically reasonable constraints are preferred. Furthermore, the ones determining the exponent in an elegant way are additionally preferable. The model presented in Chapter 4 does it. Only assuming the effort for the hearer (i.e. decoding effort) has to be minimized (plus entropy maximization), it is capable of explaining Zipf's law. If effort for the hearer and effort for the speaker are combined, $\beta \approx 2$ can be explained (Chapter 3). Inversely, when the exponent is determined, one may ask whether explaining $\beta \neq 2$ (for instance in schizophrenia or nouns) should follow from a generalization or weakening of the explanation for $\beta \approx 2$ or it requires a totally different explanation. For the former case we have that least effort for both hearer and a least effort for the speaker explains $\beta \approx 2$ and that the model in (Chapter 4) is a weakened version where scaling is explained by means of the effort for the hearer only (but retaining entropy maximization). We will thus say a model is scalable if the change from $\beta = 2$ to $\beta \neq 2$ or inversely is performed by weakening (removing constraints) or strengthening (adding constraints) it. The requirement of scalability will be labeled as *Scal.*

Model	Fin.	Word	Ref.	Presynt.	Deter.	Scal.	Type
G. K. Zipf's tautology (Zipf, 1972a; Rapoport, 1982)	√	√	√	√	√	×	C.2
Birth process (Simon, 1955)	×	√	×	√	√	×	C.1
Word length minimization (Mandelbrot, 1953; Mandelbrot, 1966; Nicolis, 1991)	×	×	×	√	×	×	A
Intermittent silence (Miller, 1957; Mandelbrot, 1953; Mandelbrot, 1966; Li, 1992)	×	×	×	√	×	×	C.1
Differential equations for Zipf's law curve (Tuldava, 1996; Montemurro, 2001; Tsonis, Schultz, and Tsonis, 1997)	×	n.a.	×	√	×	×	C.2
Random Markov process 1 (Kanter and Kessler, 1995)	√	√	×	√	√	×	C.1
Random Markov process 2 (Nicolis, 1991)	×	×	×	×	√	×	C.1
Maximum Rényi entropy (Bashkirov and Vityazev, 2000; Bashkirov, 2003)	√	√	×	n.a.	×	×	A

Maximum Tsallis entropy (symbolic dynamics under Tsallis thermodynamics (Denisov, 1997)	×	×	×	×	×	×	A
Entropy discontinuity (Harremoës and Topsøe, 2001; Harremoës and Topsøe, 2002)	×	√	×	√	√	×	B
Maximum Shannon entropy (Naranan and Balasubrahmanyan, 1992a; Naranan and Balasubrahmanyan, 1992b)	√	√	×	√	?	×	A
Maximum degenerate entropy (Naranan and Balasubrahmanyan, 1993)	√	√	×	×	×	×	A
Minimum algorithmic entropy (Balasubrahmanyan and Naranan, 1996)	√	√	×	×	√	√	A
Maximum complexity (Balasubrahmanyan and Naranan, 1996)	√	√	×	×	√	√	A
Least effort for the hearer model (Chapter 4)	√	√	√	×	×	√	A
Least effort for the hearer and the speaker model (Chapter 3)	√	√	√	×	√	√	A

Table 5.1: A summary of the traits of every model: finite lexicon (*Fin.*), word as basic unit (*Word*) word reference (*Ref.*), syntax not assumed (*Presynt.*), determined exponent (*Deter.*), scalable (*Scal.*) and type of model (*Type*). A, B, C.1, C.2 stand for optimization, stability, purely stochastic and differential equation models in column *Type*. *Fin.*, *Word* and *Ref.* are the features that a strict model need to satisfy in the context of human language. Type A models only satisfying *Fin.* and *Word* are models of Zipf' law consequences, but not models in the strict sense. *Presynt.*, *Deter.* and *Scal.* are the desired properties. *n.a.* indicates the requirement is not applicable.

The amount of assumptions made by models vary considerably. Some models assume that words arrange in strings of words thus implying the existence of syntax in various forms. For instance, some models maximize the complexity of a word sequence (Balasubrahmanyan and Naranan, 1996; Naranan and Balasubrahmanyan, 1998). Some other models imply syntax when considering the word frequency distribution in fractal binary sequences (Denisov, 1997). Assuming syntax raises the following questions: is syntax a consequence of Zipf's law or is Zipf's law a consequence of syntax? As explained in Section 5.2, modeling is not only concerned with providing explanations but also with providing the simplest explanations. If a decision has to be made between a complicated and a simple explanation has to be made, the simplest is favoured.

The model in Chapters Chapter 3-4 supports that syntax is not necessary for explaining Zipf's law in a simple way. Circular reasonings must be avoided. When dealing with type A-B models, it is easy to make the mistake of assuming the existence of an feature of language when it still does not exist. Assuming that syntax and or symbolic reference are pressures for Zipf's law can be a mistaken approach. The possibility that the crux of human language is a side-effect of a process that has nothing to do with syntax and symbolic reference can not be denied (Lieberman and Kosslyn, 2002; Gould, 1987; Pinker and Bloom, 1990). The requirement that Zipf's law is presyntactic, that is, that the explanation for Zipf's law does not assume the existence of syntax will be labeled as *Presynt.* In other words, if a model satisfies such requirement then Zipf's law is not a consequence of syntax (but can be a cause). Table 5.1 summarizes and classifies the existent Zipf's law models and summarizes the satisfaction of the previous requirements.

5.5 A null hypothesis for human language

We have seen in Section 5.2 that it is necessary to formulate appropriate null hypotheses in order to test the significance of a candidate model. Intermittent silence has been proposed for proving $\mathcal{H}_0$, but we have seen intermittent silence has nothing to do with actual word use. The requirements *Fin.* and *Word* in a biologically general context, lead to the following levels of null hypothesis. We have a set of signals $S = \{s_1, s_2, ..., s_n\}$. Assuming $P(i) = 1/n$ for the sake of simplicity we get

$$P(f) = \begin{cases} 1 & if\ f = 1/n \\ 0 & otherwise \end{cases} \tag{5.5}$$

We will call it the weak null hypothesis because whatever biased rank-frequency distribution (e.g. a power distribution) can overcome it. Zipf's law (with $\beta \approx 2$) is obviously meaningful with respect to this hypothesis.

By assuming $p(s_i) = 1/n$ for every i we are not being very precise so the weak null hypothesis can be improved. Signals refer to the objects inducing them, so there is a dependence between the frequency of a signal and the object(s) it points to. Thus, we will formulate a stronger null hypothesis assuming the definition of $p(s_i, r_j)$ given in Chapter 4. Assuming the probability that s_i and r_j are linked is $\mathcal{P}(a_{ij} = q)$, where q is a constant, we get

$$P(\mu_i = k) \sim \binom{m}{k} q^k (1-q)^{m-k}$$

where mu_i is the number of connections of the i-th signal, $q = T/(nm)$ and T is the total amount of connections. Using $p(s_i) = \mu_i/T$ (Chapter 4) we get

$$P(p(s_i) = k/T) \sim \binom{m}{k} q^k (1-q)^{m-k}$$

, Therefore, $P(f)$ is a binomial distribution, which is clearly different than a power function as Zipf's law, so the probability definitions in Chapter 4 support Zipf's law meaningfulness.

Many researchers have pointed out that human language is radically different from the communication systems found in other animals (Hauser, Chomsky, and Fitch, 2002; Deacon, 1997; Lieberman, 1991a). Syntax and symbolic reference are believed to be the distinguishing features. Nonetheless, animal communication has not been used for testing Zipf's law meaningfulness. A simple test consists of comparing the frequency distribution of the repertoire of a target non-human species with that of humans. The test has four possible configurations,

1. Zipf's law is not found for the target species.
   (a) There is no evidence of syntax and symbolic reference in the target species.
   (b) There is such evidence.
2. Zipf's law is found for the target species.
   (a) There is no evidence of syntax and symbolic reference in the target species.
   (b) If there is such evidence.

(1.a) implies the possibility that Zipf's law has to do with syntax or symbolic reference can not be denied. Zipf's law meaningfulness is supported. (1.b) implies the possibility that Zipf's law has nothing to do with syntax or symbolic reference is supported. Zipf's law meaningfulness is not supported. (2.a) favours that Zipf's law has nothing to do with syntax and symbolic reference. Zipf's law meaningfulness is not supported. (2.b) then the possibility that Zipf's law has something to do with syntax and symbolic reference is not rejected. Zipf's law meaningfulness is supported in that case.

(a) Is the general situation for non-human species with our currently available knowledge and used as a working hypothesis (Hauser, Chomsky, and Fitch, 2002). Fig. 5.1 and Fig. 5.2 show the repertoires of and captive bonobos (*Pan Paniscus*) and common ravens (*Corvus Corax*) fall into (1). Frequency distributions for the vocalizations of ravens were extracted from (Conner, 1985). As for bonobos, the analysis here is focused on the two largest repertoires: vocalizations and gestures (de Waal, 1988). As already found for the black-capped chickadee (*Parus atricapillus*) (Ficken and Ficken, 185), non-human species seem to use their units more equally frequency than expected from Zipf's law (with $\alpha \approx 1$). Syntax and symbolic reference has not been proven for those species. Therefore, such species fall into (1.a), and support Zipf's law meaningfulness. As mentioned above, restricting our analysis to the frequency distribution is a weak hypothesis test. Further work with stronger tests taking into account referents (or meanings) is necessary, along with extending the tests to other species.

5.6 Discussion

Fin. and *Word.* are necessary traits. Models not fulfilling any of the necessary traits are erroneous. A crucial requirement is *Ref.*, since reference is the major factor involved in word use. Interestingly, the only model for Zipf's law considering reference are the least effort for hearer and speaker model in Chapter 3 and the least effort for the hearer model in Chapter 4. Notice that the random Markov process in (Kanter and Kessler, 1995) takes into account that words connect through semantic constraints but it does not assume word reference. Models reproducing Zipf's law, even in a linguistically sound context, can not be tested against the null hypothesis in Section 5.5 if they do not satisfy reference. A good model not only must explain the observations under consideration but it must be testable (Popper, 1968; Medawar, 1969). Tests against additional predictions or significance tests against null hypothesis should be feasible. Therefore, *Ref.* is a linguistic requirement but also a requirement for testability. The least effort models in Chapters 3-4 are the only whose meaningfulness can be tested at this moment.

Models satisfying *Fin.* and *Word.* but not satisfying *Ref.* must be considered as models of Zipf's law consequences. Only Type A and B models can fall into such category. For type A models the consequence is some sort of optimization and for type B models some sort of stability. Models satisfying *Fin.*, *Word.* and

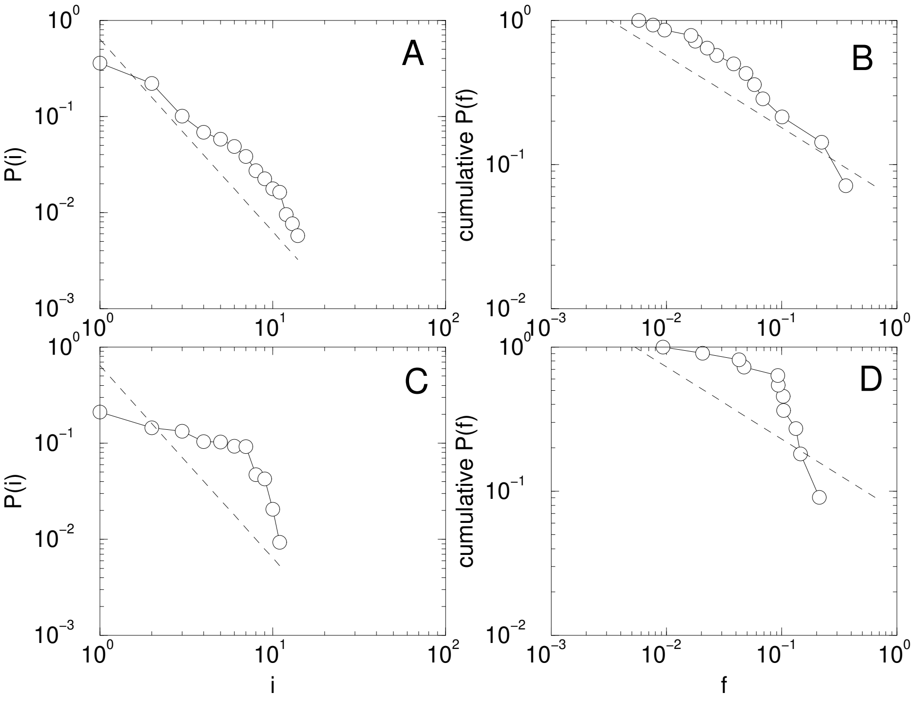

Figure 5.1: Repertoires of bonobos. A. $P(i)$, the frequency of the i-th most frequent gesture-contact pattern. B. Cumulative $P(f)$. $P(f)$ is the proportion of gesture-contact patterns whose frequency is f. C. $P(i)$, the frequency of the i-th most frequent vocal patterns. D. Cumulative $P(f)$. $P(f)$ is the proportion of vocal patterns whose frequency is f. The ideal curves for Zipf's law, that is, $P(i) = ci^{-1}$ or $P(f) = cf^{-2}$ are also shown in A-C and B-D (dashed lines), respectively.

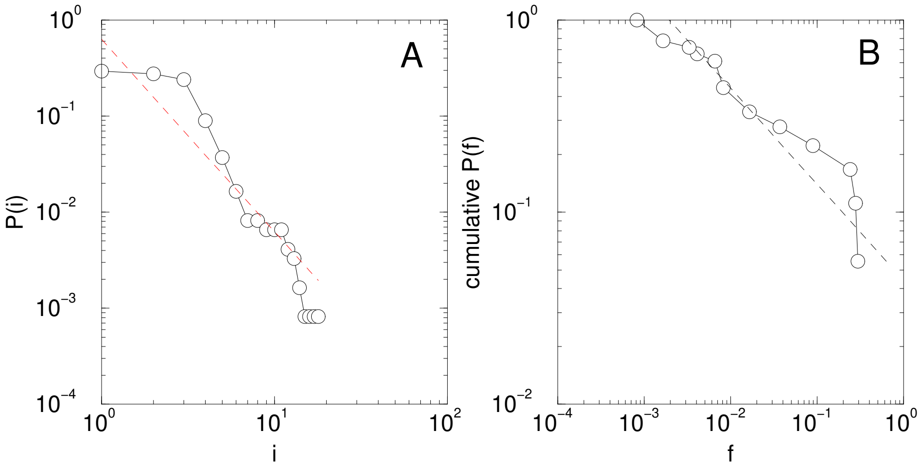

Figure 5.2: Common raven vocalizations A. $P(i)$, the frequency of the i-th most frequent vocalization. The ideal curve for Zipf's law, that is, $P(i) = ci^{-1}$, is also shown (dashed line). B. Cumulative $P(f)$. $P(f)$ is the proportion of vocalizations whose frequency is f. The ideal curve for Zipf's law, that is, $P(f) = cf^{-2}$, is also shown (dashed lines).

Ref. are strict models (except for Zipf's tautology, Table 5.1). See Table 5.2 for a classification of models according to the previous criteria. *Presynt.*, *Deter.* and *Scal.* are regarded here as desirable properties. The only model satisfying the necessary requirements, reference and the desirable properties is the dual least effort model in Chapter 3. Thereafter, the dual least effort model is not just 'an explanation among many', but the best at this moment. Nonetheless, the model has not the ultimate answer for Zipf's law. We have made a distinction between strict models of Zipf's law and models of Zipf's law consequences. The possibility that a model not based on optimization satisfying *Fin.*, *Word* and *Ref.* reproduces Zipf's law can not be denied. In a similar way, the scale-free degree distribution in complex networks can be modeled using both a purely stochastic process (Barabási and Albert, 1999) or an optimization mechanism (Chapter 9). That could turn the dual least effort model into a model of Zipf's law consequences if the alternative model is proven to be soundly defined.

Another way of classifying models for $\beta \approx 2$ is according to their amount of linguistic requirements. A good model should meet a compromise (Table 5.3). For instance, syntax (under the form of long distance correlations (Denisov, 1997) or word sequences whose complexity must be maximized (Balasubrahmanyan and Naranan, 1996) is sometimes assumed (see Table 5.1). Modeling is not only concerned about reproducing a certain pattern but also with the seek of minimal assumptions and explanations. A model with moderate requirements can help to distinguish between the origins of the regularity and its by prod-

Type of model	Models
Erroneous	Birth process (Simon, 1955). Intermittent silence (Miller, 1957; Mandelbrot, 1953; Mandelbrot, 1966; Li, 1992).
Possible models of Zipf's law consequences	Distribution differential equations (Tuldava, 1996; Montemurro, 2001; Tsonis, Schultz, and Tsonis, 1997). Maximum complexity (Balasubrahmanyan and Naranan, 1996). Entropy discontinuity (Harremoës and Topsøe, 2001; Harremoës and Topsøe, 2002). Maximum Rényi entropy (Bashkirov and Vityazev, 2000; Bashkirov, 2003).
Strict	Least effort for the hearer model (Chapter 4). Least effort for the hearer and the speaker model (Chapter 3).

Table 5.2: A classification of some models for Zipf's law in word frequencies according to the way they satisfy *Fin.*, *Word* and *Ref.*

ucts. That is why models with heavy requirements must not be rejected. Once a simple mechanism or principle has lead to Zipf's law, communication could get extra benefits. For instance, obtaining a set of word frequencies minimizing the complexity of the mechanism needed generating a sequence of words (Balasubrahmanyan and Naranan, 1996) could underly our ability to 'speak without thinking' that is evident in infants (Werker and Vouloumanos, 2001). Some models with heavy requirements are not well-defined. Such models assume that syntax results from concatenating letters that ultimately translate into sequences of words (Nicolis, 1991; Denisov, 1997). Such models neglect the existence of a mental lexicon.

Moreover, intermittent silence can not consistently explain the values of $1 < \beta < 2$ in schizophrenics where β is clearly far from 2 (Zipf, 1972a). The exponent of an intermittent silence model as in (Li, 1992) is

$$\beta_i = \frac{log|L|}{log(|L|+1)} + 1 \tag{5.6}$$

It follows from Eq. 5.6 that $1 < \beta_i < 2$ ($|L| \geq 1$ is assumed) and $\beta_i < 1.9$ implies $|L| \leq 5$. Therefore, such disease implies a repertoire of letters (or phonemes) radically smaller from that of regular speakers, which is absolutely false.

Models in Chapters 3-4 and the null hypothesis presented here suggest Zipf's law in humans has to do with communication. Nonetheless, it can not say that Zif's law implies communication (Suzuki, Tyack, and Buck, 2003). Different models arrange units according to Zipf's law without the need of reference. We have have used a priori information *i.e. 'the context is communication'* so that

Amount of requirements	Type of model
No requirements	Tautologies (e.g. (Zipf, 1972a; Tuldava, 1996; Montemurro, 2001; Tsonis, Schultz, and Tsonis, 1997)).
Little requirements	Models not determining the exponent (e.g. (Mandelbrot, 1953; Rapoport, 1982)).
Moderate requirements	Dual least effort (Chapter 3).
Heavy requirements	Models assuming syntax. (e.g. (Balasubrahmanyan and Naranan, 1996))

Table 5.3: Different amounts of requirements and some example models for Zipf's law with $\beta \approx 2$.

to reject some models. What can be done when such a priori information is missing and the only information available is a source emitting units whose frequency can computed? If such a source obeys Zipf's law with $\beta \approx 2$, we can try determine if the source codes messages in the same way humans do. First, human word frequencies are not the result of word length optimization, although word length is positively correlated with word frequency (Miller and Chomsky, 1963). The basic test for word length independence consists of restricting to words of the same length (assuming words are strings). Human language shows clearly Zipf's law (with $\beta \approx 2$) again in that case (Chapter 2), but intermittent silence (Miller and Chomsky, 1963) and the word length optimization model in (Mandelbrot, 1966) will fail. Second, the lexicon of an individual speaker is finite (and it seem to be true for the species whose communication systems have been successfully studied).

The vocabulary growth of the target source can be compared to that of a human speaker as function of text length. The vocabulary growth of a human speaker would converge to its vocabulary size (a finite amount) if text length goes to infinity. The vocabulary of the intermittent silence and Simon's birth process would go to infinity. If the probability of adding a new word in the Simon model is zero (which would make the vocabulary finite), the process fails to reproduce Zipf's law. The problem the of comparing a null hypothesis source (e.g. intermittent silence) and a target source is similar to the Turing machine halting problem in computer science (Sipser, 1999). If the target source has no bounded vocabulary size, one will have to wait till infinity to give an answer (i.e. no answer). If the target source has a finite vocabulary, we will be able to give an answer in finite time. Probably, non-human species have smaller vocabularies than humans so the time needed for finding a positive answer could be smaller than for humans. Nonetheless, applying such a kind of test to species in the wild raises many practical problems.

Here we have supported that Zipf's law has to do with communication. Assuming Zipf's law has to with communication, a further test of Zipf's law meaningfulness (in human language) concerns the complexity of the communication that Zipf's law implies. The hearer least effort model in Chapter 4 shows that if the effort for the hearer is defined as μ_i no scaling will emerge. In contrast, if it is defined as the entropy associated to the interpretation of s_i, $H(R|s_i)$, then scaling is expected ($H(R|s_i) = \log \mu_i$ (Chapter 4). As for the dual least effort model, it has been shown that defining the effort for the speaker as the lexicon size will not lead to Zipf's law, while $H(S)$, the entropy associated to signals, will (Chapter 3). To sum up, a large amount of explanations proposed for Zipf's law in linguistics and the simplicity of some of them has lead some researchers to think that Zipf's law is meaningless. But we have seen the underlying hypothesis test is often not well defined. Here we provide arguments for

- Rejecting certain existing models.
- Distinguishing between causes and consequences of Zipf's law.
- Favouring models obeying certain desirable properties
- Evaluating and classifying future models.

Researchers should have never isolated word frequencies from word meanings. By doing so, they entered the sphere where models, despite its mathematical complexity, reproduce reality just by chance. The linguistic context is relevant and thus null hypothesis must be well defined. Correctly defined null hypothesis support Zipf's law meaningfulness. This chapter puts a step forward for Zipf's law meaningfulness in a linguistic context.

Models not cited here have not been forgotten on purpose. Authors of original models not appearing in Table 5.1 are encouraged to write the present author with a reference of where the model was published, as well as a classification of the model according to Table 5.1.

Chapter 6

Syntactic dependency universals

6.1 Introduction

There is no agreement about the number of languages spoken on Earth, but estimates are in the range from $3,000$ to $10,000$ (Crystal, 1997). World languages exhibit a vast array of structural similarities and differences. Two major strategies are followed by empirically working scholars for deepening their understanding of human language faculty and language diversity. One concerned about finding *linguistic universals*, i.e. properties common to all languages. The other concerned about the differences among languages their classification, which was the former aim of typology.

The seek of linguistic universals has to face a basic problem. On the one hand, if a property is very general then it is likely to be satisfied by all languages but it is also likely to carry little information (e.g. all languages have vowels). From the other hand, if a property is more specific there is a high risk the property is only satisfied by a limited set of languages. Another general problem of most linguistic universals found (Greenberg, 1966; Greenberg, 1968; Croft, 1990) is that they are not generally portable to other disciplines or a more general framework where they can be compared and further understood. If a linguistic universal is defined in terms of the position of a certain type of word in a sentence (Greenberg, 1968), tentatively no comparison can be made with non-linguistics systems. In contrast, when studying the universal distribution of word frequencies, the so called Zipf's law for word frequencies (Zipf, 1972a), it has been hypothesized that the underlying process might be essentially the same behind solid ice melting to liquid water (Binney et al., 1992).

In this context, a recent study has shown that the presence of scaling in word frequency distributions might be a natural result of an optimization process involving a sudden phase transition (Chapter 3). In other words, scaling laws in human language would be the result of universal phenomena as those familiar to statistical physicists. Interestingly, Zipf's-law-like distributions appear in many non-linguistic domains (Ramsden and Vohradský, 1998; Furusawa and Kaneko, 2003; Burgos, 1996; Burgos and Moreno-Tovar, 1996; Balasubrahmanyan and Naranan, 2000). Specially relevant are the domains were information transfer and other linguistic metaphors are obvious, e.g. DNA sequences (Naranan and Balasubrahmanyan, 2000; Balasubrahmanyan and Naranan, 2000). Zipf's law is portable.

Common language universals are just empirical generalizations resulting from inductive studies. Linguists in that field are well aware of this fact, and the fact that universals have to be explained themselves (which means they are not laws but general observations together with inductively formulated hypotheses). In contrast, physics has its own understanding of universal laws, where the macroscopic regularities of different systems are explained with the same basic mechanism. Critical systems, for example, are grouped into universality classes that only depend on dimensionality, symmetry of the order parameter and symmetry and range of interactions (Chaikin and Lubensky, 1995; Stanley et al., 1996; Stanley et al., 2000). Statistical physics thus provides a well defined understanding of universality that is largely system independent. This is certainly not the common view in linguistics (Croft, 1990). Empirical evidence supports the possibility that a large number of systems arising in disparate disciplines such as physics, biology and economics might share some key properties involving their large-scale organization (Stanley et al., 2000). One of the most remarkable of these universal laws is related to scale invariance, i. e. the presence of a hierarchical organization that repeats itself at very different scales. Using tools from statistical physics we present evidence for previously unreported syntactic universals that are both portable and tentatively enough specific.

Most of linguistic research is done in the domain of descriptive approaches (e.g Chomsky's standard and extended 'theories' (Uriagereka, 1998)). But descriptions are not explanations. Linguistic explanation is not possible without the construction of a linguistic theory containing universal language laws (Köhler, 1987). The search for language and text laws in the sense of the philosophy of science - and not in the improper sense commonly used in mainstream linguistics - is strongly connected with the work of Gabriel Altmann and his school (Altmann, 1978; Altmann, 1993). Since 1983, 'Synergetic linguistics' has been systematically developed by Köhler (Köhler, 1987) as a first linguistic theory on the basis of the central axiom of language as a self-organizing system and on functional explanation. This theory was first presented in the field of the lexicon (Köhler, 1986) and later extended to morphology and syntax (Köhler, 1999; Köhler and Altmann, 2000).

The aim of the present chapter is investigating potential syntactic universals, which in turn may provide clues for understanding the origins of language. Since syntax is a crucial feature in human language uniqueness (Hauser, Chomsky, and Fitch, 2002; Lieberman, 1991a), we will focus on syntactic universals. Different non-excluding positions are taken for explaining linguistic universals. To cite some examples, an underlying universal grammar (Uriagereka, 1998), genetic encoding (Pinker and Bloom, 1990; Pinker, 1996) or functional constraints (Hawkins, 1994; Hawkins, 1992; Lieberman, 1991b). Syntax involves a set of rules for combining words into phrases and sentences. Such rules ultimately define *explicit* syntactic relations among words that can be directly mapped into a graph capturing most of the global features of the underlying rules. Such a network-based approach has provided new insights into semantic webs (Steyvers and Tenenbaum, 2001; Sigman and Cecchi, 2002; Motter et al., 2002; Kinouchi et al., 2002). Capturing global syntactic information using a network has been attempted. The global structure of word interactions in short contexts in sentences has been studied (Ferrer i Cancho and V. Solé, 2001; Dorogovtsev and Mendes, 2001). Although about 87% of syntactic relationships take place at distance lower or equal than 2 (Chapter 8), such early work lacks both a linguistically precise definition of link and fails in capturing the characteristic long-distance correlations of words in sentences (Chomsky, 1957). The proportion of incorrect syntactic dependency links captured with a window of length 2 as in (Ferrer i Cancho and V. Solé, 2001) is

$$\epsilon_2 = \frac{(n-1)(1-p_1) + (n-2)(1-p_2)}{2n-3}$$

where n is the length of the sentence and p_1 and p_2 are, respectively, the probability that two words at distance 1 and 2 are syntactically linked. When $n \to \infty$ we have

$$\epsilon_2 = 1 - \frac{p_1 + p_2}{2}.$$

Knowing $p_1 = 0.70$ and $p_2 = 0.17$ (Chapter 8) we get

$$\epsilon_2 = 0.56.$$

That is, one half of links are syntactically meaningless. Using a window of length 1 we have

$$\epsilon_1 = \frac{(n-1)(1-p_1)}{n-1}.$$

When $n \to \infty$ we get $\epsilon_1 = 1 - p_1$, which gives $\epsilon_1 = 0.30$, which is still high. A precise definition of syntactic link is thus required. Here, we study the architecture of syntactic graphs and show that they display small world patterns, scale free structure, a well-defined hierarchical organization and assortative mixing (Barabási and Albert, 2002; Dorogovtsev and Mendes, 2002; Newman, 2003b). Three different European languages will be used. The chapter is organized as follows. The three datasets are presented together with a brief definition of the procedure used for building the networks in Section 6.2. The key measures used in this study are presented 6.3, with the basic results reported in section 6.4. A comparison between sentence-level patterns and global patterns is presented in 6.5. A general discussion and summary is given in Section 6.6.

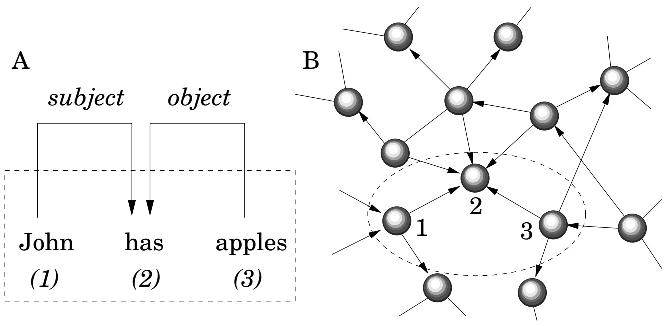

Figure 6.1: A. The syntactic structure of a simple sentence. Here words define the nodes in a graph and the binary relations (arcs) represent syntactic dependencies. Here we assume arcs go from a modifier to its head. The proper noun 'John' and the verb 'has' are syntactically dependent in the sentence. 'John' is a modifier of the verb 'has', which is its head. Similarly, the action of 'has' is modified by its object 'apples'. B. Mapping the syntactic dependency structure of the sentence in A into a global syntactic dependency network.

6.2 The syntactic dependency network

The networks that are analyzed here have been defined according to the dependency grammar formalism. Dependency grammar is a family of grammatical formalisms (Melčuk, 1988; Hudson, 1984; Sleator and Temperley, 1991), which share the assumption that syntactic structure consists of lexical nodes (representing words) and binary relations (dependencies) linking them. This formalism thus naturally defines a network structure. In this approximation, a dependency relation connects a pair of words. Most of links are directed and the arc usually goes from the head word to its modifier. Head and modifier are primitive concepts in the dependency grammar formalism (Fig. 6.1 A). In some cases, such as coordination, there is no clear direction (Melčuk, 2002). Since that cases are rather uncommon, we will assume that links in the datasets used here have a direction and assign an arbitrary direction to the undirected cases. Syntactic relations are thus binary, usually directed and sometimes typed in order to distinguish different kinds of dependency.

We define a syntactic dependency network as a set of n words $V = \{s_i\}, (i = 1, ..., n)$ and an adjacency matrix $A = \{a_{ij}\}$. s_i can be a modifier word of the head s_j in a sentence if $a_{ij} = 1$ ($a_{ij} = 0$ otherwise). Here, we assume arcs go from a modifier to its head. The syntactic dependency structure of a sentence can be seen as a subset of all possible syntactic links contained on a global network (Fig. 6.1 B). More precisely, the structure of a sentence is a subgraph (a tree) of the global network that is induced by the words in the sentence (Bollobás, 1998).

Different measures can be defined on A allowing to test the presence of certain interesting features such as the small-world effect (Watts and Strogatz, 1998) and scale invariance (Barabási and Albert, 1999). Such measures can also be used for finding similarities and differences among different networks (see Section III).

The common formal property of dependency representations (compared to other syntactic representations) is the lack of explicit encoding for phrases as in the phrase structure formalism (Chomsky, 1957) and later developments (Uriagereka, 1998). Dependency grammar regards phrases as emergent patterns of syntactic dependency interactions. Statistical studies about phrase-structure-based grammars have been performed and reveal that the properties of syntactic constructs map to only a few distributions (Köhler, 1999; Köhler and Altmann, 2000), suggesting a reduced set of principles behind syntactic structures.

We studied three global syntactic dependencies networks from three European languages: Czech, German and Romanian. Because of the reduced availability of data, the language set is unintentionally restricted to the Slavic, Germanic and Italic families. These languages are not intended to be representative of every family. We are not taking the common inductive position in the study of linguistic universals. Chapter 7 gives further support for our position. We mention the families these languages belong to in order to show how distant these languages are. Probably not enough distant for standard methods in linguistics for defining universals but enough distant for our concerns here. Syntactic dependency networks were build collecting all words and syntactic dependency links appearing in three corpora (a corpus is a collection of sentences). Here, $a_{ij} = 1$ if an arc from the i-th word to the j-th word has appeared in a sentence at least once and $a_{ij} = 0$ otherwise. Punctuation marks and loops (arcs from a word to itself) were rejected in all three corpora. The study was performed on the largest connected component of the networks. Sentences with less than two words were rejected.

The corpora analyzed here are a Cezch corpus by Ludmila Uhlířová and Jan Králík, the Dependency Grammar Annotator Corpus for Romanian and the Negra Corpus for German (Appendix A). The German corpus is the most sparse of them. It is important to notice that while the missing links in the German corpus obey no clear regularity, links in the Czech corpus are mostly function words, specially prepositions, the annotators did not link because they treated them as grammatical markers. The links that are missing are those corresponding to the most connected words types in the remaining corpora.

6.3 Network properties

In order to properly look for syntactic dependency patterns, we need to consider several statistical measures mainly based on the undirected version of the network for simplicity reasons. These measures allow to categorize networks in terms of:

1. *Small world structure.* Two key quantities allow to characterize the global organization of a complex network. The first is the so called *average path*

lenght D, defined as $D = \langle D_{min}(i,j)\rangle$, where $\langle ...\rangle$ is the average operator over all pairs (s_i, s_j) in the network, where $D_{min}(i,j)$ indicates the length of the shortest path between nodes i and j. D was calculated on the largest connected component of the networks. The second measure is C, the so called clustering coefficient, defined as the probability that two vertices (e.g. words) that are neighbors of a given vertex are neighbors of each other. C is defined as $\langle C_i\rangle$ where $\langle ...\rangle$ is the average over all vertices and C_i, the clustering coefficient of the i-th vertex, is easily defined from the adjacency matrix as

$$C_i = \frac{2}{k_i(k_i-1)}\sum_{j=1}^{n} a_{ij}\left(\sum_{l=j+1} a_{il}a_{jl}\right) \tag{6.1}$$

where k_i is the degree of the i-th vertex. Erdös-Rényi graphs have a binomial degree distribution that can be approximated by a Poissonian distribution (Barabási and Albert, 2002; Dorogovtsev and Mendes, 2002; Newman, 2003b). An Erdös-Rényi graph is such that the probability that two vertices are linked is the same for all different pairs of vertices. Erdös-Rényi graphs with an average degree $\langle k\rangle$ are such that $C_{random} \approx \langle k\rangle/(n-1)$ and the path length follows (Newman, 2000)

$$D_{random} \approx \frac{\log n}{\log \langle k\rangle .} \tag{6.2}$$

It is said that a network exhibits the small-world phenomenon when $D \approx D_{random}$ (Watts and Strogatz, 1998). The key difference between an Erdös-Rényi graph and a real network is often $C \gg C_{random}$ (Barabási and Albert, 2002; Dorogovtsev and Mendes, 2002; Newman, 2003b).

2. *Heterogeneity.* A different type of characterization of the statistical properties of a complex network is given by the degree distribution $P(k)$. Although the degree distribution of Erdös-Rényi graphs is Poisson, most complex networks are actually characterized by highly heterogeneous distributions, i.e. they can be described by a degree distribution $P(k) \sim k^{-\gamma}\phi(k/k_c)$, where $\phi(k/k_c)$ introduces a cut-off at some characteristic scale k_c. The simplest test of scale invariance is thus performed by looking at $P(k)$, the probability that a vertex has degree k, often obeying (Barabási and Albert, 2002; Dorogovtsev and Mendes, 2002; Newman, 2003b)

$$P(k) \sim k^{-\gamma}.$$

The degree distribution is the only statistical measure where link direction will be considered. Therefore, input and output degree will be also analyzed.

3. *Hierarchical organization.* Some scaling properties indicate the presence of hierarchical organization and modularity in complex networks. When

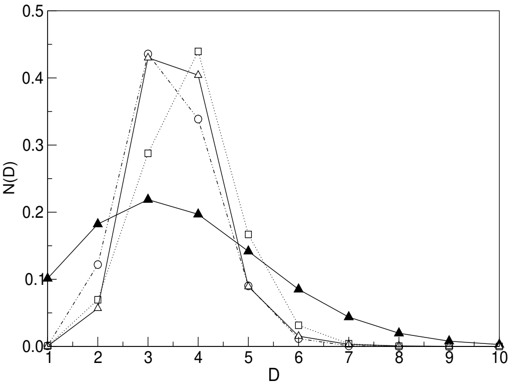

Figure 6.2: Shortest path length distributions for the three syntactic networks analyzed here. The symbols correspond to: Romanian (circles), Czech (triangles) and German (squares) respectively. The three distributions are peaked around an average distance of $D \approx 3.5$ degrees of separation. The expected distribution for a Poissonian graph is also shown (filled triangles), using the same average distance.

studying $C(k)$, i. e. the clustering coefficient as a function of the degree k, certain networks have been shown to behave as (Ravasz et al., 2002; Ravasz and Barabási, 2002)

$$C(k) \sim k^{-\theta} \tag{6.3}$$

with $\theta \approx 1$ (Ravasz et al., 2002). Hierarchical patterns are specially important here, since tree-like structures derived from the analysis of sentence structure strongly claim for a hierarchy.

4. *Betweenness centrality.* While many real networks exhibit scaling in their degree distributions, the value of the exponent γ is not universal, the betweenness centrality distribution is less varying (Goh et al., 2002) although it fails to work as a network classification method (Barthélemy, 2003). The betweenness centrality of a vertex v, $g(v)$, is a measure of the number of minimum distance paths running through v, that is defined as (Goh et al., 2002)

$$g(v) = \sum_{i \neq j} \frac{G_v(i,j)}{G(i,j)}$$

where $G_v(i,j)$ is the number of shortest pathways between i and j running through v and $G(i,j) = \sum_v G_v(i,j)$. Many real networks obey

$$P(g) \sim g^{-\eta}$$

where $P(g)$ is the proportion of vertices whose betweenness centrality is g. The betweenness centrality was calculated using Brandes' algorithm (Brandes, 2001).

5. *Assortativeness.*

   A network is said to show *assortative mixing* if the nodes in the network that have many connections tend to be connected to other nodes with many connections A network is said to show disassortative mixing if the highly connected nodes tend to be connected to nodes with few connections. The Pearson correlation coefficient Γ defined in (Newman, 2002) measures the type of mixing with $\Gamma > 0$ for assortative mixing and $\Gamma < 0$ for disassortative mixing. Such correlation function can be defined as

$$\Gamma = \frac{c\sum_i j_i k_i - [c\sum_i \frac{1}{2}(j_i + k_i))]^2}{c\sum_i \frac{1}{2}(j_i^2 + k_i^2) - [c\sum_i \frac{1}{2}(j_i + k_i))]^2} \tag{6.4}$$

   where j_i and k_i are the degrees of the vertices at the ends of the i-th edge, with $i = 1...m$, $c = 1/m$ and m being the number of edges. Disassortative mixing ($\Gamma < 0$) is shared by Internet, World-Wide Web, protein interactions, neural networks and food webs. In contrast, different kinds of social relationships are assortative ($\Gamma > 0$) (Newman, 2002; Newman, 2003a).

6.4 Results

The first relevant result of our study is the presence of small world structure in the syntax graph. As shown by our analysis (see Table 6.1 for a summary), syntactic networks show $D \approx 3.5$ degrees of separation. The values of D and C are very similar for Czech and Romanian. A certain degree of variation for German can be attributed to the fact it is the most sparse dataset. Thus, D is overestimated and C is underestimated. Nonetheless, all networks have D close to D_{random} which is the the hallmark of the small-world phenomenon (Watts and Strogatz, 1998). The fact that $C \gg C_{random}$ indicates (Table 6.1) that the organization of syntactic networks strongly differs from the Erdös-Rényi graphs. Additionally, we have also studied the frequency of short path lengths for the three networks. As shown in Figure 6.2, the three distributions are actually very similar, thus suggesting a common pattern of organization. When we compare the observed distributions to the expectation from a random Poissonian graph (indicated by filled triangles), they strongly differ. Although the average value is the same, syntactic networks are much more narrowly distributed. This was early observed in the analysis of World Wide Web (Adamic, 1999).

The second result concerns the presence of scaling in their degree distributions. The scaling exponents are summarized in Table 6.1. For the undirected graph, we have found that the networks are scale free with $\gamma \approx 2.2$. Additionally, Fig. 6.3 shows $P(k)$ for input and output degrees (see Table 6.1 for the specific values observed). With the exception of the Czech corpus, they display well-defined scale-free distributions. The Czech data set departs from the power

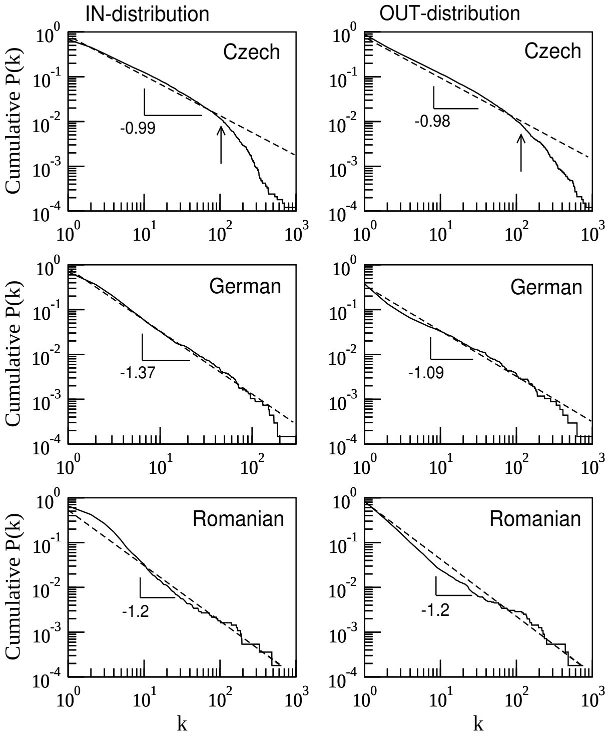

Figure 6.3: Left. Cumulative degree distributions for the three corpora. Here the proportion of vertices whose input and output degree is k are shown. The plots are computed using the cumulative distributions $P_{\geq}(k) = \sum_{j \geq k} P(j)$. The arrows in the plots on top indicate the deviation from the scaling behavior in the Czech corpus.

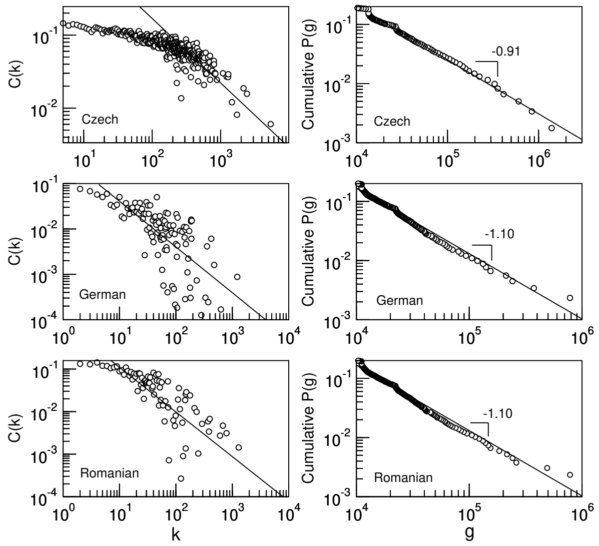

Figure 6.4: Left: $C(k)$, the clustering coefficient versus degree k for the for the three corpora. In all three pictures the scaling relation $C(k) \sim k^{-1}$ is shown for comparison. Right: the corresponding (cumulative) $P(g)$, the proportion of vertices whose betweenness centrality is g.

law for $k > 10^2$. Thus highly connected words appear underestimated in this case, consistently with the limitations of this corpus discussed in section 6.2. These power laws fully confirm the presence of scaling at all levels of language organization (Hřebíček, 1995).

Complex networks display hierarchical structure (Ravasz et al., 2002). Fig. 6.4 (left column) shows the distribution of clustering coefficients $C(k)$ against degree for the different corpora. We observe skewed distributions of $C(k)$ (which are not power laws), as in other systems displaying hierarchical organization, such as the World Wide Web (see Fig. 3(c) in (Ravasz and Barabási, 2002)).

In order to measure to what extent word syntactic dependency degree k is related to word frequency, f, we calculated the average value of f versus k (6.5) and found a power distribution of the form

$$f \sim k^{\zeta} \tag{6.5}$$

where $\zeta \approx 1$ (Table 6.1) indicates a linear relationship (Fig. 6.5). The higher values of ζ for German can be attributed to the sparseness of the German corpus.

Highly connected words tend to be not interconnected among them. Since degree and frequency are positively correlated (Eq. 6.5 and Fig. 6.5) one easily concludes, as a visual examination will reveal, that the most connected words are function words (i.e. prepositions, articles, determiners,...). Disassortative mixing ($\Gamma < 0$) tells us that function words tend to avoid linking each other.

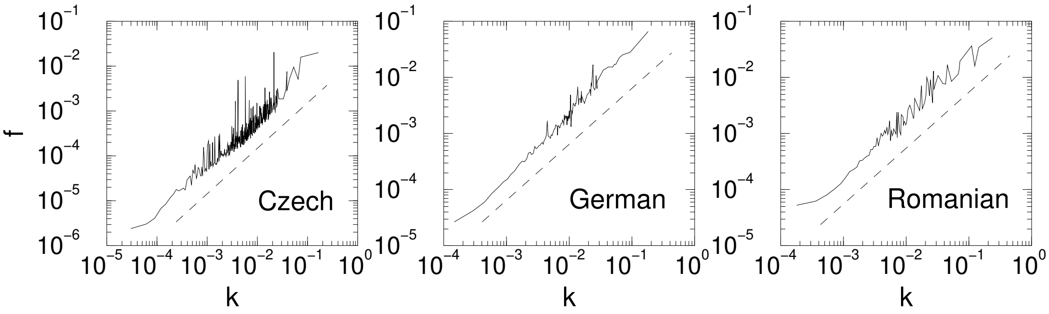

Figure 6.5: Average word frequency f of words having degree k. Dashed lines indicate the slope of $f \sim k$, in agreement with real series.

This consistently explains why the Czech corpus has a value of Γ clearly greater than that of the remaining languages. We already mentioned in section II that most of the missing links in the Czech corpus are those involving function words such as prepositions, which are in turn the words responsible for a tendency to avoid links among highly connected words. Γ is thus overestimated in the Czech network.

The scaling exponent γ is somewhat variable, but the scaling exponents obtained for the betweenness centrality measure are more narrowly constrained (table 6.1). Although again the Czech corpus deviates from the other two (in an expected way) the two other corpora display a remarkable similarity. $P(g)$ distribution, with $\eta = 2.1$. Is is worth mentioning that the fits are very accurate and give an exponent that seems to be different from those reported in most complex networks analyzed so far, typically $\eta \in [2.0, 2.2]$ (Goh et al., 2002). The behavior of $P(g)$ in Fig. 6.4 with a domain with scaling with $\eta \approx 2.1$ for German and Romanian suggests a common pattern is shared. The deviation of Cezch from the remaining networks may be explained by its lack of hub words.

The behavior of $C(k)$ (Fig. 6.4, left) differs from the independence of the vertex degree found in Poisson networks and certain scale-free network models (Ravasz and Barabási, 2002). Such behavior $C(k)$ is also different from Eq. 6.3 with $\theta = 1$ that is clearly found in synonymy networks and suggested in actor networks (Ravasz et al., 2002) and metabolic networks (Ravasz et al., 2002). In contrast, such behavior is similar to that of the World Wide Web and Internet at the Autonomous System level (Ravasz and Barabási, 2002). The similar shape of $C(k)$ in the three syntactic dependency networks suggests all languages belong to the same universality class.

Besides word co-occurrence networks and the syntactic dependency networks presented here, other types of linguistic networks have been studied. Networks were nodes are words or concepts and links and semantic relations are known to show $C \gg C_{random}$ with $d \approx d_{random}$ and power distribution of degrees with and exponent $\gamma \in [3, 3.5]$. For the Roget's Thesaurus, assortative mixing ($\Gamma =$

	Czech	German	Romanian	Software graph	Proteome [1]
n	33336	6789	5563	1993	1846
$< k >$	13.4	4.6	5.1	5.0	2.0
C	0.1	0.02	0.09	0.17	2.2×10^{-2}
C_{random}	$4 \cdot 10^{-4}$	$6 \cdot 10^{-6}$	$9.2 \cdot 10^{-4}$	2×10^{-3}	1.5×10^{-3}
D	3.5	3.8	3.4	4.85	7.14
D_{random}	4	5.7	5.2	4.72	9.0
Γ	-0.06	-0.18	-0.2	-0.08	-0.16
γ	2.29 ± 0.09	2.23 ± 0.02	2.19 ± 0.02	2.85 ± 0.11	2.5 ($k_c \sim 20$)
γ_{in}	1.99 ± 0.01	2.37 ± 0.02	2.2 ± 0.01	-	-
γ_{out}	1.98 ± 0.01	2.09 ± 0.01	2.2 ± 0.01	-	-
η	1.91 ± 0.007	2.1 ± 0.005	2.1 ± 0.005	2.0	2.2
θ	Skewed	Skewed	Skewed	Skewed	1.0
ζ	1.03 ± 0.02	1.18 ± 0.01	1.06 ± 0.02	-	-

Table 6.1: A summary of the basic features that characterize the potential universal features exhibited by the three syntactic dependency networks analyzed here. n is the number of vertices of the networks, $< k >$ is the average degree, C is the clustering coefficient, C_{random} is the value of C of an Erdös-Rényi network. D is the average minimum vertex-vertex distance, D_{random} is the value of D for an Erdös-Rényi graph. Γ is the Pearson correlation coefficient. γ, γ_{in} and γ_{out} are respectively, the exponents of the undirected degree distribution, input degree distribution, output degree distribution. η, θ and ζ are, respectively, the exponents of the betweenness centrality distribution, the clustering versus degree and the frequency versus degree. Two further examples of complex networks are shown. One is a technological graph (a software network analyzed in (Valverde, Ferrer i Cancho, and Solé, 2002)) and the second is a biological web: the protein interaction map of yeast (Jeong et al., 2001). Here *skewed* indicates that the distribution $C(k)$ decays with k but not necessarily following a power law.

0.157) is found (Newman, 2003b). (Steyvers and Tenenbaum, 2001; Sigman and Cecchi, 2002; Motter et al., 2002; Kinouchi et al., 2002). In contrast, syntactic dependency networks have $\gamma \in [2.11, 2.29]$ and disassortative mixing (Table 6.1), suggesting networks of semantic relations have exponents belonging to a different universality class. Further work, including more precise measures, such as the exponent of $P(g)$, should be carried out for semantic networks.

6.5 Global versus sentence-level patterns

We have mentioned that there is a high risk that very general linguistic universals carry no information. Similarly, one may argue that the regularities

	Czech	Romanian	German
d_{global}	$2.3 \cdot 10^{-4}$	$1.3 \cdot 10^{-3}$	$1.2 \cdot 10^{-3}$
$< d_{sentence} >$	0.88	0.75	0.83
C_{global}	0.1	0.09	0.02
$< C_{sentence} >$	0	0	0
Γ_{global}	-0.06	-0.2	-0.18
$< \Gamma_{sentence} >$	-0.4	-0.51	-0.64

Table 6.2: Summary of global versus sentence network traits. d_{global}, C_{global} and Γ_{global} are, respectively, the normalized average vertex-vertex distance, the clustering coefficient and the Pearson correlation coefficient of a given global syntactic dependency network. $d_{sentence}$, $C_{sentence}$ and $\Gamma_{sentence}$ are, respectively, the normalized average vertex-vertex distance, the clustering coefficient and the Pearson correlation coefficient of a given sentence syntactic dependency network. $\langle x \rangle$ stands for the average value of the variable x over all sentence syntactic dependency networks where x is defined.

encountered here are not significant unless it is shown they are not a trivial consequence of some pattern already present in the syntactic structure of isolated sentences. In order to dismiss such possibility, we define d_{global} and $d_{sentence}$ as the normalized vertex-vertex distance of the global dependency networks and a sentence dependency network. The normalized average vertex-vertex distance is defined here as

$$d = \frac{D-1}{D_{max}-1}$$

where $D_{max} = \frac{n+1}{3}$, the maximum distance of a connected network with n nodes (Chapter 9). Similarly, we define C_{global} and $C_{sentence}$ for the clustering coefficient and Γ_{global} and $\Gamma_{sentence}$ for the Pearson correlation coefficient. The clustering coefficient of whatever syntactic dependency structure is $C_{sentence} = 0$, since the syntactic dependency structure is defined with no cycles (Melčuk, 1988). We find $C_{global} \gg C_{sentence}$ and $d_{global} \ll d_{sentence}$ (Table 6.2). $\Gamma_{sentence}$ is clearly different than Γ_{global}, although disassortative mixing is found in both cases.

Besides, one may think that the global degree distribution is scale-free because the degree distribution of the syntactic dependency structure of a sentence is already scale free. $P_{sentence}(k)$, the probability that the degree of a word of a word in a sentence is k is not a power function of k (Fig. 6.6). Actually, the data point suggests an exponential fit. To sum up, we conclude that scaling in $P(k)$, small-world with significantly high C and the proper value of γ are features emerging at the macroscopic scale. The global patterns discussed above are emergent features that show up at the global level.

6.6 Discussion

We have presented a study of the statistical patterns of organization displayed by three different corpus in this chapter. The study reveals that, as it occurs at other levels of language organization (Steyvers and Tenenbaum, 2001; Sigman and Cecchi, 2002; Motter et al., 2002; Kinouchi et al., 2002), scaling is widespread. The analysis shows that syntax is a small world and displays a well defined and potentially universal global structure. These features can be properly quantified and have been shown to be rather homogeneous. No one can speak of a linguistic universal in standard linguistics before hundreds of languages have been investigated according to a sophisticated system of criteria for the selection of the languages to study in order to cover any known type of language family, etc. in a balanced proportion. Here, the context is different. Here, we have the backup of universality in physics. Different patterns of statistical patterns of complex networks are the result of very general mechanisms. To cite some examples, the preferential attachment principle generates scale-free networks (Barabási and Albert, 1999; Dorogovtsev and Mendes, 2003) as well as a conflict between vertex-vertex distance and link density minimization (Chapter 9). Randomness in the way vertices are linked is a source of small-worldness (Watts and Strogatz, 1998). 'How universal small-world is in syntactic dependency networks?' is a question more related to how randomly are words linked than to how much sufficient is the language sample examined here. This allows us to conclude that a new class of *potential language universals* can be defined on quantitative grounds. We have taken the position of formulating the most specific hypotheses according to out currently available data. Our findings do not exclude investigating more languages in order to validate our hypotheses.

Understanding the origins of syntax implies understanding what is essential in human language. Recent studies have explored this question by using mathematical models inspired in evolutionary dynamics (Nowak and Krakauer, 1999; Nowak, Plotkin, and Jansen, 2000; Nowak, 2000b). However, the study of the origins of language is usually dissociated from the quantitative analysis of real syntactic structures. General statistical regularities that human language obeys at different scales are known (Hřebíček, 1995; Köhler, 1999; Köhler and Altmann, 2000). The statistical pattern reported here could serve as validation of existent formal approaches to the origins of syntax. What is reported here is specially suitable for recent evolutionary approaches to the origins of language (Nowak and Krakauer, 1999; Nowak, Plotkin, and Jansen, 2000; Nowak, 2000b), since they reduce syntax to word pairwise relationships.

Linguists can decide not to consider certain word types as vertices in the syntactic dependency structure. For instance, annotators in the Czech corpus decided that prepositions are not vertices. That way, we have seen that different statistical regularities are distorted, e.g. disassortative mixing almost disappears and degree distributions are truncated with regard to the remaining corpora. If the degree distribution is truncated, describing degree distributions requires more complex functions. If simplicity is a desirable property, syntactic descriptions should consider prepositions and similar word types as words in the strict sense. Annotators should be aware of the consequences of their decision about the local structure of sentences wit regard to global statistical patterns.

Syntactic dependency networks do not imply recursion, that is regarded as a crucial trait of the language faculty (Hauser, Chomsky, and Fitch, 2002). Nonetheless, different non-trivial traits that recursion needs have been quantified quantified:

- Disassortative mixing tells us that labour is divided in human language. Linking words tend to avoid connections among them.

- Hierarchical organization tells us that syntactic dependency networks not only define the syntactically correct links (if certain context freedom is assumed) but also a top-down hierarchical organization that is the basis of phrase structure formalisms such as X-bar (Bickerton, 1990).

- Small-worldness is a necessary condition for recursion. If mental navigation (Kinouchi et al., 2002) in the syntactic dependency structure can not be performed reasonably fast, recursion can not take place. Pressures for fast vocal communication are known to exist (Lieberman, 1991b; Hawkins, 1992).

An interesting prospect of our work is that explaining certain linguistic universals may also explain other network patterns outside the linguistic context without loss of generality. We have seen for the non-trivial properties analyzed here that human languages are likely to belong to the same universality class. Such a class is a novel way of understanding world languages internal coherence and essential similarity. In contrast, when regarding other systems, human languages exhibit both unique and matching features.

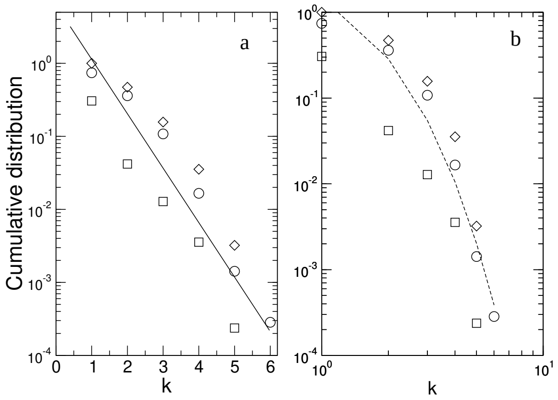

Figure 6.6: Cumulative $P_{sentence}(k)$ for Czech (circles), German (squares) and Romanian (diamonds). Here linear-log (a) and log-log (b) plots have been used, indicating an exponential-like decay. $P_{sentence}(k)$ is the probability that a word has degree k in the syntactic dependency structure of a sentence. Notice that $P_{\geq}(1)$ is less than 1 for Czech and German since the sentence dependency trees are not complete. If $P_{sentence}$ was a power function, a straight line should appear in log-log scale. The German corpus is so sparse than its appearance is dubious. Statistics are shown for L^* the typical sentence length. We have $L^* = 12$ for Czech and German and $L^* = 6$ for Romanian.

Chapter 7

From referential principles to language

7.1 Introduction

Although many species possess rudimentary communication systems (Hauser, 1996; Ujhelyi, 1996), humans seem to be unique with regard to making use of syntax (Hauser, Chomsky, and Fitch, 2002) and symbolic reference (Deacon, 1997; Donald, 1991; Donald, 1998). Recent approaches to the evolution of language formalize *why* syntax is selectively advantageous compared to isolated signal communication systems (Nowak and Krakauer, 1999; Nowak, Plotkin, and Jansen, 2000), but they do not explain *how* signals naturally combine. We have seen in Chapter 3 that if a communication system minimizes both the effort of the speaker and that of the hearer, signal frequencies will be distributed according to Zipf's law (with $\beta \approx 2$). Here we will show that such a communication principle gives rise not only to signals that have many traits in common with the linking words in real human languages, but also to a rudimentary sort of syntax and symbolic reference. Furthermore, we will identify different statistical patterns found in real syntactic dependency networks (Chapter 6). Finding Zipf's law in an animal communication system will be shown to be sufficient condition for such a rudimentary form of language.

We have seen Zipf's law meaningfulness is supported in the context of human language (Chapter 5). Zipf's law (with $\beta \approx 2$) is obtained when simultaneously minimizing the communicative effort of the speaker and the hearer, the former needing to minimize the uncertainty associated to the selection of a word and the latter needing to minimize the uncertainty in the interpretation of the meaning of a word (Chapter 3). Here we take Zipf's law for granted and examine its predictions.

7.2 From Zipf's law to language

We assume a general communication framework, and thus define a set of signals $S = \{s_1, ..., s_i, ..., s_n\}$ and a set of objects of reference $R = \{r_1, .., r_j, ..., r_m\}$. We define a matrix of signal-object associations $A = \{a_{ij}\}$ $(1 \leq i \leq n \ , \ 1 \leq j \leq m)$ where $a_{ij} = 1$ if the i-th signal and the j-th object are associated and $a_{ij} = 0$ otherwise. The matrix A defines a bipartite graph $G_{n,m}$ (Bollobás, 1998) with edges corresponding to the 1s in A. Let us write p_k for the proportion of signals with k links. Assuming that all objects have a similar frequency, the relative frequency of a signal is naturally defined as proportional to the number of objects it is connected to, so Eq. 1.1 becomes

$$p_k \sim k^{-\beta}. \tag{7.1}$$

Here we shall only assume Zipf's law, or rather Eq. 7.1. Our model for $G_{n,m}$ will be as follows: given the numbers n and m of signals and objects, and for each k the proportion p_k of signals connected to k objects, the graph $G_{n,m}$ is chosen uniformly at random from among all bipartite graphs with these properties. Equivalently, having decided the degree $d(s_i)$, i.e., the number of associated objects, of each signal appropriately, we join s_i to a random set of $d(s_i)$ objects, independently of the other signals. We investigate properties that the resulting graph has with high probability, noting that any such property is a very natural consequence of Zipf's law. Note that there is a transition in the model at $\beta = 2$, due to the rapid change in the number of edges as β is varied about this value. More precisely, the average degree of a signal is $\sum_{k=1}^{m} kp_k$. The infinite form of this sum converges if and only if $\beta > 2$; in this range the average degree is asymptotically constant as m increases. In contrast, for $\beta = 2$ the average degree grows logarithmically with m and, for $\beta < 2$, as a power of m. In asymptotic analysis we shall thus consider $\beta = 2 + \epsilon$ for some small ϵ.

Different theoretical approaches to syntax assume that a connection between a pair of syntactically linked words implies that the words are semantically compatible (Chomsky, 1965a; Helbig, 1992). Here we assume that a pair of signals are connected to each other through a common object, which, acting as a rudimentary meaning, defines the semantic compatibility of such a pair. Thus, given the signal-object graph $G_{n,m}$, we define a signal-signal graph G_n whose vertices are the signals s_i, in which two signals are joined if in $G_{n,m}$ they are joined to one or more common objects. For various reasons, our grammar is not a grammar in the strict sense, but rather a protogrammar, from which full human language can easily evolve. First, notice that such a grammar lacks word order (Sleator and Temperley, 1991) and link direction (Melčuk, 1989). Second, such a grammar does not imply (but allows) recursion (Hauser, Chomsky, and Fitch, 2002). Syntax likewise involves overcoming the limits of memory for keeping track of the complex relationships between words within the same sentence (Lieberman, 1991a). A phrase can be formed by choosing a pair of words (u, v) in G_n and all the words in a path from u to v. Total freedom for forming phrases only exists when there is a path between every pair of vertices, that is, when the network is connected.

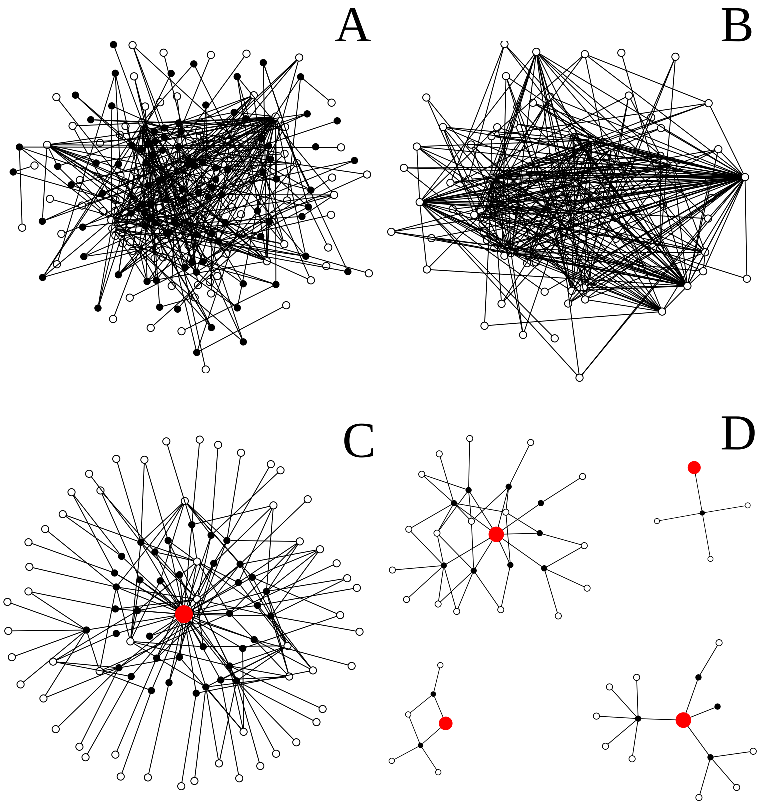

Figure 7.1: Examples of $G_{n,m}$ (A) and G_n (B) for $\beta = 2$ and $n = m = 100$. White and black circles are signals and objects, respectively. First and second neighbors of the most connected signal (red circle) in A (C). This and other highly connected signals are the forerunners of linking words (e.g. prepositions and conjunctions) in human language. First and second neighbors of other signals (red circles) in A (D). Linkers in human language have have (a) poor (or absent) referential power (Givón, 2002), (b) high frequency (Baayen, 2001) and (c) many connections with referentially powerful words (more precisely, disassortative mixing; Chapter 6). Highly connected signals satisfy (a) since the uncertainty associated to the interpretation of a signal grows with its number of links (Fig. 7.1 C). Satisfying (b) follows trivially from the proportionality relationship between frequency and number of objects. (c) follows from the skewed and long-tailed distributions for p_k and q_k. Disassortative mixing supports (b-c) (Fig. 7.4).

When $\beta = 2 + \epsilon$, with high probability G_n is almost connected in the sense that almost all signals lie in a single component (the limiting proportion does not tend to zero as $\epsilon \to 0$). See Fig. 7.1 A-B and Fig. 7.2 A. Almost connectedness is easy to derive mathematically, although there is no space here for the details. There are two key requirements. Firstly, the 'expected neighbourhood expansion factor' f must be greater than one. Roughly speaking, f is the average number of nodes (here signals) within distance $\ell + 1$ of a given node s, divided by the number within distance ℓ, for ℓ in a suitable range. If ℓ is neither too small nor too large and t is a node at distance ℓ from s, then the expected number of neighbours of t at distance $\ell + 1$ from s is essentially independent of ℓ. Here, noting that one 'step' in G_n corresponds to two in $G_{n,m}$, one can check that

$$f = \frac{n}{m} \sum_k (k-1) k p_k.$$

For $m = n$ this is greater than 1 for $\beta < 3.54..$, and in particular for $\beta \approx 2$. Given that $f > 1$, standard methods show that there will be a single giant component, and that all other signals are in 'small' components with only a few vertices. In fact, that is true for $m \ll n \log n$. For $\beta = 2 + \epsilon$, one can easily check that asymptotically order $c(\epsilon)n$ signals are in small components, and the rest of G_n is connected. Here $c(\epsilon)$ is a constant depending on ϵ and approaching zero as $\epsilon \to 0$. More precisely, this is true for $m \ll n/\epsilon$.

7.3 From Zipf's law to syntactic dependency universals

We will show that different statistical patterns in real syntactic dependency networks (Chapter 6) can also be found in G_n. We start with the degree distribution, defining q_k as the proportion of signals having degree k in G_n, recalling that two signals are joined in G_n if they are associated with at least one common object. Let Z be the degree in G_n of a random signal s_i, so $q_k = \Pr(Z = k)$. With $\beta = 2 + \epsilon$ it is very unlikely that two given signals are joined to two or more common objects, so Z is essentially

$$\sum_{r_j \sim s_i} d(r_j) - 1,$$

where $d(r_j)$ is the degree in $G_{n,m}$ (number of associated signals) of an object r_j, and the sum is over all objects associated to s_i. Now, as whether a signal other than s_i is associated to r_j is independent of $s_i \sim r_j$, the terms $d(r_j) - 1$ in the sum behave like essentially independent Poisson distributions, each with mean $\lambda = (n/m) \sum_k k p_k$, which tends to a constant as $n, m \to \infty$ with n/m constant. The distribution of Z does not have a very simple form, but its tail does: the sum of Poisson distributions is again Poisson, and is very unlikely to exceed its mean, here $\lambda d(s_i)$, by any given factor when the mean is large. Thus,

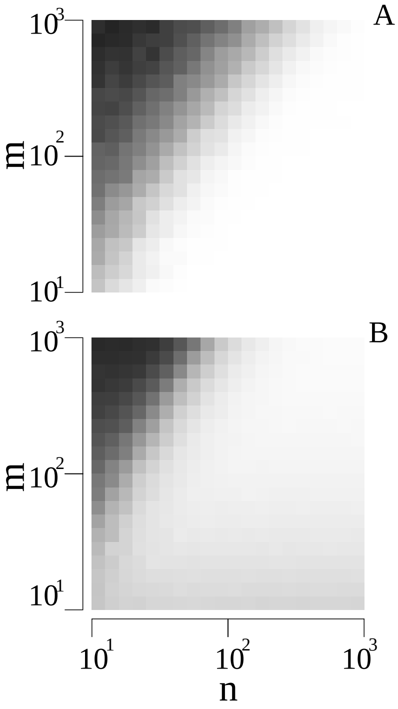

Figure 7.2: Proportion of vertices in the largest connected component of G_n versus n and m. A gray scale from 0 (black) to 1 (white) is used. A. Signal degrees in $G_{n,m}$ following a power distribution with $\beta = 2$. B. Signal degrees in $G_{n,m}$ following a binomial distribution with the same expected degree as in B. All values were calculated using numerical estimations for $10 \leq n, m < 10^3$. Loops in G_n are forbidden.

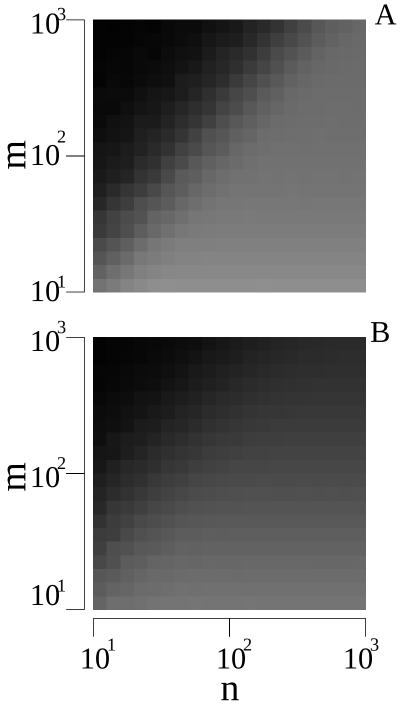

Figure 7.3: C, the clustering coefficient in G_n versus n and m. A gray scale from 0 (back) to 1 (white) is used. A. Signal degrees in $G_{n,m}$ following a power distribution with $\beta = 2$. B. Signal degrees in $G_{n,m}$ following a binomial distribution with the same expected degree as in B. All values were calculated using numerical estimations for $10 \leq n, m < 10^3$. Loops in G_n are forbidden.

one can check that as $k \to \infty$ (keeping n/m fixed) we have

$$q_k \sim ck^{-\beta}, \tag{7.2}$$

with c a positive constant. Thus, while the exact distribution of Z is not a power law, Z does have a power-law tail, with the same exponent β as the signal degrees and the signal frequency distribution (Eq. 1.1). Eq. 7.2 is consistent with the analysis of real syntactic dependency networks, where the proportion of words having k syntactic links with other words is $\sim k^{-\gamma}$ with $\gamma \approx 2.2$ (Chapter 6). Note that γ is in turn close to the the typical Zipf's law exponent.

Syntactic theory regards certain function words such as prepositions and conjunctions as linkers (Melčuk, 1989), that is words serving for combining words for forming complex sentences. The most connected signals in $G_{n,m}$ share many features with real linking words (Fig. 7.1).

Assuming Zipf's law with $\beta \approx 2$, other patterns of real syntactic dependency networks (Chapter 6) are recovered

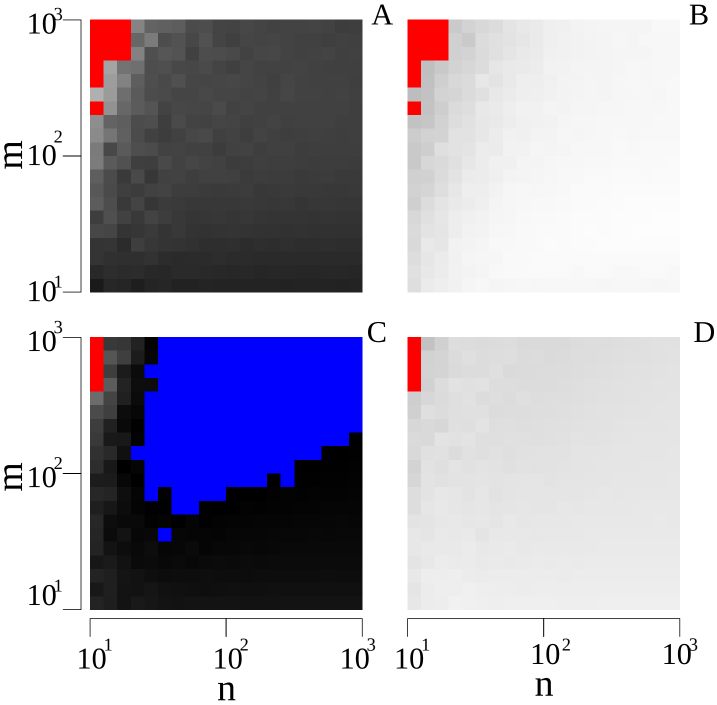

Figure 7.4: $|\Gamma_{degree}|$ and $1 - |\Delta\Gamma_{degree}|/2$ in G_n versus n and m. Γ_{degree} is the normalized correlation of the degrees at either ends of an edge. $\Delta\Gamma_{degree} = \Gamma_{degree}^{real} - \Gamma_{degree}$. $\Gamma_{degree}^{real} \approx -0.2$ for human language (Chapter 6). Undefined, positive and negative values of Γ_{degree} and $\Delta\Gamma_{degree}$ appear in red, blue and gray scale, respectively. A gray scale from 0 (black) to 1 (white) is used. A. Γ_{degree} when signal degrees in $G_{n,m}$ follow a power distribution with $\beta = 2$. B. The same as in A. for $\Delta\Gamma_{degree}$. C. Γ_{degree} when signal degrees in $G_{n,m}$ follow a binomial distribution with the same expected degree as in A. D. The same as in C. for $\Delta\Gamma_{degree}$. All values were calculated using numerical estimations for $10 \leq n, m < 10^3$. Loops in G_n are forbidden.

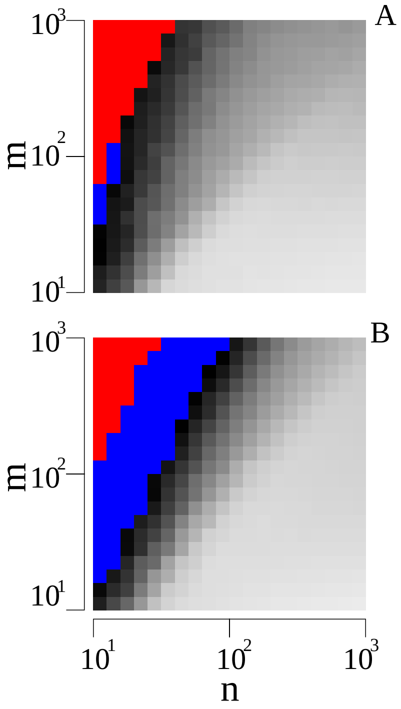

Figure 7.5: $|\Gamma_{clustering}|$ in G_n versus n and m. $\Gamma_{clustering}$ is the correlation between $C(k)$, the clustering of a vertex whose degree is k, and k. Undefined, positive and negative values of $\Gamma_{clustering}$ appear in red, blue and gray scale, respectively. A gray scale from 0 (black) to 1 (white) is used. A. Signal degrees in $G_{n,m}$ following a power distribution with $\beta = 2$. B. Signal degrees in $G_{n,m}$ following a binomial distribution with the same expected degree as in B. All values were calculated using numerical estimations for $10 \leq n, m < 10^3$. Loops in G_n are forbidden.

Predictions of $p_k \sim k^{-\beta}$ with $\beta \approx 2$	
Pattern	Origin
Connectedness	$\langle k \rangle_P$
Significantly high clustering	$\langle k \rangle_P$
Hierarchical organization	$\langle k \rangle_P$
Disassortative mixing	P
Signal degree distribution in G_n ($q_k \sim k^{-\gamma}$ with $\gamma \approx \beta$)	P
Linking words as in humans (referentially useless highly connected words)	P

Table 7.1: Predictions of Zipf's law with $\beta \approx 2$ and the origin of the predictions. We distinguish two types of origin: $\langle k \rangle_P$ if the prediction basically depends on the expected signal degree in $G_{n,m}$ or P if the prediction can not be explained by the previous one.

- High clustering (Watts and Strogatz, 1998) in Fig. 7.3.

- Disassortative mixing (Newman, 2002) in Fig 7.4.

- Hierarchical organization (Ravasz and Barabási, 2002) in Fig 7.5.

Such patterns are clear for $n \sim m$, where connectedness is warranted and disappear when $m \gg n$. In order to understand the nature of the predictions made by Zipf's law, we define a null hypothesis. We have a distribution for signal degrees $P = \{p_1, ..., p_k, ..., p_m\}$ and distribution for object degrees $\Pi = \{\pi_1, ..., \pi_k, ..., \pi_n\}$. We assume Π is fixed and defined for simplicity as $\pi \sim binomial(n, \langle k \rangle_P / m)$ where $\langle k \rangle_P$ is the expectation operator over the distribution P. We replace P by P' for building the null hypothesis. We define $P' = \{p'_k\}$ where $p'_k \sim binomial(m, \langle k \rangle_P / n)$. Using P', it can be seen that qualitatively similar results are obtained for clustering (Fig. 7.3), hierarchical organization (Fig. 7.5) but not for disassortative mixing (Fig. 7.4). The order of the largest connected component behaves qualitatively in the same way for P and P' (Fig. 7.2). Thus, some patterns are qualitatively caused by a certain average degree and not a specific distribution. The only patterns that depend on the degree distribution are the type of mixing and obviously q_k, the degree distribution in G_n and the presence of linking words as in human language (referentially useless highly connected words). We conclude that what was found for only three languages in Chapter 6 is qualitatively universal, since Zipf's law is universal. Table 7.1 summarizes the universal predictions that Zipfs's law makes and the origins of the prediction, that is, P or $\langle k \rangle_P$.

7.4 From Zipf's law to symbolic reference

We assume Deacon's understanding of symbolic reference (Deacon, 1997), which is in turn mostly based on Peirce's (Peirce, 1932). According to Peirce classification, reference is iconic when the mapping between signals and objects is made trough physical similarity. Reference is indexical when the mapping between signals and objects is made trough temporal or spatial correlation. Convention is the way signals and objects get linked in symbolic reference. The present thesis operates at high level of abstraction. A binary matrix $A = \{a_{ij}\}$ tells which objects every signal refers two. There is no constraint on the type of references A defines. Nonetheless, for the purpose of the present discussion, it is important to consider associations in A as non-symbolic, because we want to explain how a symbolic system emerges from A. The definition of a symbol is a yet open and highly debatable issue (Cangelosi, Greco, and Harnad, 2002).. General overview of theories of reference are found in (Sinha, 1999; Cangelosi, Greco, and Harnad, 2002).

The section is organized as follows. Section 7.4.1 defines a higher form of reference on $G_{n,m}$. Section 7.4.2 shows how such higher form of reference captures Deacon's understanding of symbolic reference. Section 7.4.3 explains how such a higher form of symbolic reference overcomes all the criticisms to Deacon's view.

7.4.1 A higher form of indexical reference

The configuration of $G_{n,m}$ can lead to higher order forms reference. If s_i and r_j are linked and r_j and s_k are also linked ($i \neq k$), then a signal-signal referential association between s_i and s_k is formed via r_j in only two steps. If the network is connected, that is, there is a path between every pair of vertices, then reference between every signal and another signal in the network is allowed via less than $2(n-1)$ links. The word 'allowed' is important here. Such a higher form of reference needs connectedness, but connectedness does not imply such a higher form of reference is fully developed. Singnal-object connections define allowed reference links among signals and objects. An objective measure of the effective referential power of the communication system is not given by the amount of links, but by $I(R, S)$, the information transfer (Shannon, 1948). Certain signal-object configurations that are connected destroyed reference. The situation is well illustrated by an example. Chapter 4 showed that a complete bipartite graph (every signal connected to very object) has no referential power at all, that is $I(R, S) = 0$. Such a higher order form of reference is only effective when connectedness is achieved by maintaining a reasonably high values of $I(R, S)$. We do not want a too restrictive definition for our higher form of indexical reference, so we will focus on almost connectedness instead of full connectedness. We will say a network is almost connected when the largest connected component contains more than about 1/2 of the nodes. To sum up, our higher form of indexical reference is defined as (at least) almost connectedness with reasonably high values of $I(R, S)$. For the sake of simplicity, using 'connectedness' alone

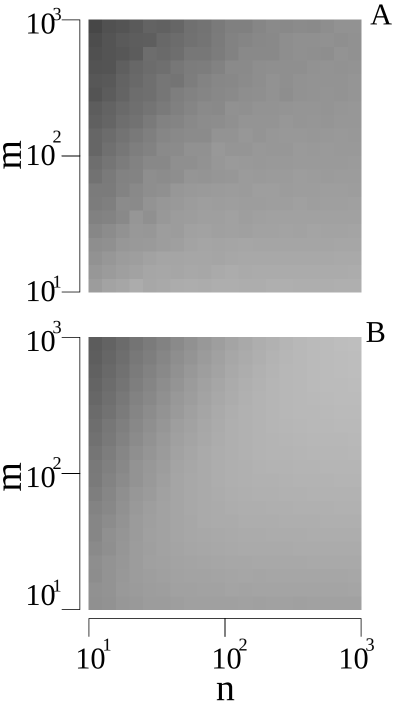

Figure 7.6: $\bar{I}(R,S)$, the normalized information transfer in $G_{n,m}$ versus n and m. A gray scale from 0 (black) to 1 (white) is used. A. Signal degrees in $G_{n,m}$ following a power distribution with $\beta = 2$. B. Signal degrees in $G_{n,m}$ following a binomial distribution with the same expected degree as in B. All values were calculated using numerical estimations for $10 \leq n, m < 10^3$. Loops in G_n are forbidden.

will always imply almost connectedness in what follows.

Assuming Π is fixed, the value of $I(R,S)$ depends on the distribution P. Now we show that Zipf's law complies with the previous definition of a higher order form of reference. The present chapter relies on the assuming that the proportion of signals with k obeys Eq. 7.1 with $\beta = 2$ for Zipf's law. We define the normalized information transfer as

$$\bar{I}(R,S) = I(R,S)/\log m$$

where $\log m$ is the maximum value of $H(R)$. Fig. 7.2 shows that G_n and therefore $G_{n,m}$ is almost connected. Reasonably high values of $\bar{I}(R,S)$ are obtained assuming Zipf's law (Fig. 7.6 A). Using P' instead of P for signal degrees, similar results are obtained in G_n (Fig. 7.6 B). In both cases, $m \gg n$ must be avoided. Therefore, the mean signal degree Zipf's law provides with such a higher form of reference.

7.4.2 Deacon's view

Our higher order form of reference shares several traits with Deacon's understanding of symbolic reference (Deacon, 1997):

- Binary relationships between signals. Deacon's view is lucid in identifying the role of combinatorics as the essence of symbolic reference: *Symbolic reference derives from combinatorial possibilities and impossibilities, and we therefore depend on combinations both to discover it (during learning) and to make use of it (during communication).* (Deacon, 1997)

- Signal-signal referential associations (Deacon, 1997, p. 83).

- Symbolic reference needs word ambiguity. A communication system maximizing the information transfer (i.e. minimizing the effort for the hearer) by mapping every object with a distinctive signal (Chapter 3), is not expected to be connected at all.

- It integrates syntax and symbolic reference. In Deacon's words *'Thus, syntax structure is an integral feature of symbolic reference, not something added and separate'* (Deacon, 1997, p. 100). We have seen in Chapter 7 that connectedness in G_n gives rise to a system with different common traits with human syntax. Connectedness in G_n follows trivially from connectedness in $G_{n,m}$.

- The symbolic threshold that Deacon hypothesizes for the origins of symbolic system (Deacon, 1997, p.79) can be explained better knowing that transitions to connectedness are a sudden phenomenon. More precisely, Zipf's law, a particular instance of such higher order type of reference, could have been a sudden event once a certain threshold in the balance between hearer and speaker needs is crossed (Chapter 3).

7.4.3 Criticisms to Deacon's view

Criticism to Deacon's view of language can be followed in a series of papers (Poeppel, 1997; Hurford, 1998; Hudson, 1999b). Now, we list and discuss the different criticisms concerning Deacon's interpretation of symbolic reference:

- *The clarity of the definition of symbolic reference.*
  Jim Hurford states in his review that Deacon's explanations are not unequivocally clear to linguists and philosophers (Hurford, 1998). Let us cite Peirce's definition of symbolic reference (borrowed from (Oliphant, 2002)): a symbol is *'a sign (a signal here), which refers to the object it denotes by virtue of law, usually an association of general ideas, which operates to cause the symbol to be interpreted as referring to that object (Peirce, 1932, page 276)'*. We will argue that standard approaches to symbolic reference are not an example of clarity either. Here, the proper interpretation of 'law' is that the association between signal and object is not governed by

any of the mechanisms of icons and indices. That does not exclude that a symbol can behave as an index or an icon under certain circumstances. It is important to notice that Peirce's definition is the same as saying: a symbol is something that is not either an index or an icon'. Peirce's definition is via negation. It is not constructive. Therefore, Peirce sweeps all the problem of defining a symbol under the magic word *law* or *convention.* The original definition is not clear and Deacon tries to go beyond Peirce's definition in a constructive way. Even more constructive, because of its precision, is the definition of a higher form of reference given in Section 7.4.1.

- *What makes a difference between index and symbol.*
  Hudson finds strange that Deacon points out that the difference between a symbol and an index is that a symbol is embedded in a system that connects it to other symbols (Hudson, 1999b, p.3). Without connectedness, our higher order reference consistent with Deacon's view disappears. Therefore, if connections with other symbols disappear (which is more general than connectedness), then symbolic reference disappears. Connections are crucial in Deacon's view. Our definition, constitent with Deacon's view,

- *Confusion between syntagmatic or paradigmatic relations.*
  Hudson (Hudson, 1999b, p.3) points out that Deacon is not clear when specifying the kinds of inter-symbol relations he has in mind, syntagmatic or paradigmantic. It seems Hudson is not catching that Deacons is intentionally using both. The type of reference presented in Section 7.4.1 and the definition of syntax given in Section 7.3 are indissociable phenomena. Syntactic links relationships emerge from referential connections.

- *The misuse of the term reference.*
  In our view, Deacon provides lucid new insights into the understanding of symbol. He makes a serious effort trying to unveil the relationship between symbolic reference and indexical reference. Both forms are different but not dissociated. Symbolic reference in humans is embedded on an indexical reference system. As Deacon's puts it, *'Words point to objects (reference) and words point to other words (sense), but we use the sense to pick out the reference, not vice versa (Deacon, 1997, p. 83).* He argues there is a referential relationship between words which is regarded by Hurford as a misunderstanding since *reference for a philosopher, or for a linguist, is a relation between an element in a language, like the word John and something in the world (its referent), such the flesh-and-blood person John* (Hurford, 1998). We believe Hurford's criticism needs to be reconsidered for the following reasons:

  - The path to symbolic reference used here reconciles last Hurford's criticisms and Deacon's signal-signal associations because we assume signals connect through common objects. There is no need to violate the orthodox definition of reference. A link between signals s_i and s_j with $i \neq j$ (a link in G_n) implies two proper definitions of reference (two links in $G_{n,m}$): one from s_i to some r_k and another from such r_k to s_j, provided that r_k is linked to both s_i and s_j.

  - It is a tight understanding of indexical reference: the basis of indexical reference are spatial or temporal correlations between signals and objects of reference. Association is a basic task that our brain performs. Why should not be possible signal-signal associations? Why can not we call them references? Hurford seems to be trapped by a dualistic division between elements in a language and elements in the world. Once a word is uttered, it becomes an element in the world. An uttered word and the noise of a fridge are elements *in the world.* Eventually, words are both linguistic and world elements.
  - It raises a methodological problem: it leaves no room for abstraction which is one the basic tools of science.

- *What makes a symbolic system complex.*
  Hudson (Hudson, 1999b, p.3) says that *'A symbol system must, by definition, involve some minimum degree of complexity'* contradicting Deacon's argument that human uniqueness lies the use of symbols, not in complexity. Probably, Hudson is taking a reductionist position. The higher order reference based on connectedness plus constraints is an emergent phenomenon. What make a reference system symbolic is not a new type of signal-association but the emergence of connectedness.

- *How a single symbol could had got off the ground.*
  Hudson (Hudson, 1999b, p.3) argues there can never be a first symbol because symbols (in Deacons's view), by matter of definition, are embedded in a system. Hudson invokes gradual evolution while an entire symbolic system could have stemmed from simple disconnected referential system. Chapter 3 points out that a perfect communication system (that is a disconnected communication system) is hard to maintain when the number of objects to describe goes to infinity. Probably, as suggested in Chapter 3 two steps are needed. One for setting a perfect communication system and another for moving the system to the transition where Zipf's law is found. We may not be able to start symbolic reference from scratch. In that sense, Hudson comment is sharp.

- *The dominance of symbolic reference over syntax.*
  The review by Poeppel does not poise any serious question about the Deacon's interpretation of symbolic reference, but is skeptical about the dominance of symbolic reference over syntax that Deacon defends. Poeppel's criticism has no consequence in the present thesis, because we give the same weight to syntax and symbolic reference. It is important that Deacons' view is not reductionist but integrative although he makes emphasis on symbolics reference: *'Thus syntactic structure is an integral feature*

> *of symbolic reference, not something added and separate* (Deacon, 1997, p. 100). While researchers are divided when considering syntax (Hauser, Chomsky, and Fitch, 2002) or symbolic reference as the essence of human language (Deacon, 1997; Donald, 1991; Donald, 1998) we hypothesize that syntax and symbolic reference are two sides of the same coin, i.e., connectedness in signal-signal associations. The transition to syntax and symbolic reference would have been as abrupt as the transition to Zipf's law (Chapter 3). The reader should not believe that syntax and symbolic reference are therefore the same. One thing is syntax, combining semantically compatible signals, and another is a higher order form of reference with regard to indexical reference, where a new form of reference appears when a referential link between a signal and an object propagates to another signal linked to the same object. Symbolic reference implies such higher order reference.

Hudson (Hudson, 1999b, p.5) proposes a mechanism by which children learn a symbols and their meanings in way that he believes to solve the dark points of Deacon's view. The procedure follows three stages,

1. Learn the word as an index for a co-present object X.

2. Learn the word as an index for a co-present object X to which the speaker is paying attention.

3. Learn the word as an index for an object X that it is not co-present but which the speaker is paying attention to.

The previous procedure works well for specific words but not for very ambiguous words such as 'get'. Hudson illustrates his procedure with the word 'cat'. This is a very favourable case, since quantitative measures show that nouns are on average less ambiguous than average words (Chapter 4). The previous procedure can not be used for function words. If a symbol is whatever that can be learned using Hudson procedure, then function words are not symbols. For instance, the word 'of' can not be learned as an index. Furthermore, the word 'have' is both a function word (auxiliary verb) and a content word (e.g. meaning possession) Only the content word can be learned by Hudson's procedure. So, it follows that 'have' is both a symbol and not a symbol at the same time, which is a contradiction.

7.4.4 Discussion

The definition of syntax given in Chapter 7.3 equates connectedness and syntax. The previous definition is only valid in a reductionist context where syntax is dissociated from other linguistic dimensions. If language is basically defined as reference plus syntax, connectedness provides syntax but not reference. We mentioned there is a debate about what is more essential for human language, syntax or symbolic reference. Here, we understand both are indissociable traits. Nonetheless, its reasonable to think that symbolic reference preceded syntax in the very origins of language (at least for a short period of time). This is supported by the definition of syntactic signal-signal connections that stem from signal-signal referential connections.

Although the definition of a symbol is a highly debatable issue (Cangelosi, Greco, and Harnad, 2002), there is relative consensus about certain questions:

- At least in Peirce's view, symbols are linked to its referent by convention (Deacon, 1997; Sinha, 1999; Cangelosi, Greco, and Harnad, 2002; Oliphant, 2002).
- Symbols rest on non-symbolic types of reference (indexical relations of co-occurrence (Hudson, 1999b; Sinha, 1999) or categorial representations in (Harnad, 1990)).

The last point of consensus is specially important, because our higher order reference can be constructed from any type of reference but keeping the original type of reference. Our definition just needs the trivial requirement that the source reference systems is non-symbolic, so it is consistent with the with the consensus about symbolic reference.

We can replace the negative definition of symbolic reference by a positive definition and shed light on the magic word 'convention' or 'law' used in Peirce's definition. With connectedness on mind, it is easy to see that the referents of a signal propagate to other signals, the second neighbors of the signal in $G_{n,m}$ in such a way that the interpretation of a signal s_i is allowed to be not only a function of the objects s_i is linked to but also a function but also a function of the signals sharing objects with s_i. In the end, that should be true for 4th, 6th, 6th, ..., $2x$-th neighbours, but it is reasonable to think that the weight of the contribution of such neighbours should decay with x. The fact that words are disambiguated using different types of context (e.g. other words in the same sentence) that have not necessarily nothing to do with 'element in the word' (as Hurford would say) supports the idea that ambiguities are solved via neighbouring signals and objects. Such disambiguation via internal elements can only take place under connectedness in $G_{n.m}$ and is the reason by which symbols can be learned and also the reason by which index-based word learning procedure fails..

Given the similarities between human words and the higher form of reference presented here, its easy to suggest that symbolic reference needs such form of reference. The type of reference formally presented here makes interesting predictions:

- The transition to symbolic reference in preceded by a period of stasis in the ontogeny and probably in the phylogeny of human language (Vihman and Depaolis, 2000) because transitions to connectedness are usually sharp phenomena (Bollobás, 2001).
- If a communication system is not symbolic (or equivalently it has no syntax) then it is not organized according to Zipf's law. The previous prediction is supported by non-human species frequency distributions in Chapter 5.

The previous strong predictions suggest that the type of reference presented here is actually a sufficient for symbolic reference. Future research should make emphasis on determining if such form of reference is actually a sufficient condition.

The present work clearly puts a step forward by providing the first formal definition of a type of reference that could be basically equivalent to symbolic reference. We have seen that Zipf's law with $\beta \approx 2$ is a sufficient condition for connectedness and the existence of linking words. Connectedness is in turn a necessary condition for syntax with recursion and symbolic reference in signal-object associations. Therefore, we can determine if a species has some sort of language without having deciphered its utterances. Evidence of Zipf's law (with $\beta \approx 2$) and reference are enough. A Rosetta Stone is not needed. Zipf's law is not the hallmark of human language (Chapter 5) but the hallmark of conflicting communication constraints solving (Chapter 3). While the presence of Zipf's law in cetaceans and primates is still an open problem (McCowan, Hanser, and Doyle, 1999; McCowan, Doyle, and Hanser, 2002; Janik, 1999), the possibility that other species have converged to Zipf's law, and thus to rudimentary form of language, can not be denied.

Chapter 8

Euclidean distance minimization

8.1 Introduction

A key trait of the faculty of language-narrow sense is a mechanism combining a finite set of words for yielding a potentially infinite amount of sentences (von Humboldt, 1972; Hauser, Chomsky, and Fitch, 2002). This capacity yields the so-called discrete infinity. World languages exhibit many common traits, the so-called linguistic universals (Greenberg, 1968; Dryer, 1989). Here some of them are examined. There are many constraints limiting the usage of discrete infinity. Lung capacity imposes limits on the length of actual spoken sentences, whereas working memory imposes limits on the complexity of sentences if they are to be understandable (Hauser, Chomsky, and Fitch, 2002). The fact that about 70% of the links in sentences are formed between words at distance 1 and 17% are formed at distance 2 in the Dependency Grammar Annotator Corpus (Appendix A). suggests some sort of Euclidean distance minimization principle.

It is generally assumed that the universal properties of languages are ultimately explained by both biological (i.e. innate) and functional factors (Hawkins, 1992). The distance between syntactically related items in sentences is a basic ingredient of the cost of a sentence (Gibson, 2000; Hawkins, 1994) and has been used for explaining word order universals (Hawkins, 1994). Cost minimization, or equivalently, least effort principles are a successful explanation for other universals in quantitative linguistics. For instance, Zipf's law (Zipf, 1972a) for word frequencies can be explained by minimizing hearer and speaker communicative needs (Chapter 3). Here it is shown that non-crossing dependency networks follow from such principles, which leads to other successful predictions.

Minimizing the sum of the distances between linked vertices on a network where vertices follow a sequence is known as the minimum linear arrangement (m.l.a.) problem (Díaz, Petit, and Serna, 2002). The fact that human utterances are linear (i.e. a row of basic units) was early emphasized by the French linguist

Ferdinand de Saussure (de Saussure, 1916). In present-day words, utterances have only one dimension. Sentences are thus good candidates for optimization principles operating in the way words are arranged.

Suppose we have a network whose set of vertices is V and its set of arcs is A (a directed graph). Suppose $\pi(v)$ is the position of vertex v . Then, $d(u, v) = |\pi(u) - \pi(v)|$ is the distance between vertices u and v (where $u, v \in V$). The m.l.a. problem consists of finding the π such that $\Omega(\pi, A) = \sum_{(u,v)\in A} d(u, v)$ is minimum.

Taking the words in a sentence as vertices and arcs as syntactic dependencies, the remaining of the chapter is devoted to understand that Ω is actually minimized in sentences and to explain its consequences linguistic universals and the first language spoken on Earth.

8.2 Methods

Two different sources of data were used for the present study. Both are collections of sentences with its syntactic dependency structure. The first is the Dependency Grammar Annotator corpus and the second is a Czech corpus by Ludmila Uhlířová and Jan Králík (Appendix A). When having complete structures was critical, only the Romanian corpus was used. Punctuation marks were absent so distances between words are true distances in both cases. Czech and Romanian are both SVO languages.

Random (undirected) trees were generated from scratch for different purposes. The procedure was the following:

1. Start with a network with n vertices and no edges.

2. Choose a pair of vertices (i.e. words) chosen at random (all words have the same probability to be chosen).

3. Link them if the pair of words is not linked and the graph is kept without cycles.

4. If the the network has less than $n - 1$ links go to 2.

5. End.

A fast heuristic algorithm for solving the m.l.a. problem (Koren and Harel, 2002) is used for simplicity. Finding the m.l.a on a generic graph is a very hard computational problem (Díaz, Petit, and Serna, 2002; Garey and Johnson, 1979). If the network is a tree exact computationally affordable algorithms exist (Shiloach, 1979; Chung, 1984). Results will show that the exact approaches are not needed in this context. Numerical calculations up to $n = 11$ showed that the algorithm in (Koren and Harel, 2002) always finds the optimum on trees.

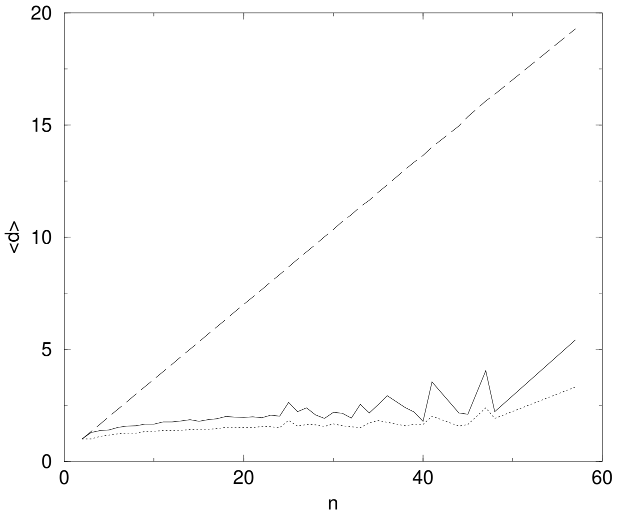

Figure 8.1: The average value of $< d >$, the mean edge length, versus the length of the sentence, n, for real (solid line) and optimized (dotted line) syntactic dependency structures. A control $< d >$ was calculated by scrambling the words in every sentence 1000 times and averaging $< d >$ (long dashed). The latter case is $< d >= \frac{n+1}{3}$ as expected.

8.3 Results

We define the average value of d, the distance between linked vertices, defined as

$$< d >= \frac{1}{n-1}\Omega(\pi, A)$$

where n is the length of the sentence (notice $|A| = n-1$). Fig. 8.1 shows $< d >$ as a function of the n for real Czech and Romanian sentences. A control series is calculated scrambling position of vertices (while the network structure remains the same) and calculating Ω again (Ω is used instead of $\Omega(\pi, A)$ for brevity). It follows for the latter case that

$$P(d) = \frac{2(n-d)}{n(n-1)} \tag{8.1}$$

Replacing the previous equation into

$$E[d] = \sum_{d=1} dP(d)$$

(where E is the expectation operator) we get

$$E[d] = \frac{n+1}{3} \tag{8.2}$$

after some algebra. It becomes evident that real sentences minimize Ω far from the upper bound provided by Eq. 8.2. The fact that $< d >$ for real sentences is greater than that of the heuristic approximation shows that using the exact algorithm for trees (Shiloach, 1979; Chung, 1984) is not necessary in this context.

Fig. 8.2 shows the amount of edge crossings, C, as a function of the tree size, n, for random trees and their m.l.a. approximations. It can be seen that the amount of links is very small for the m.l.a. counterparts. The fact is supported by theorems stating, under rather general conditions, that for whatever crossing arrangement π on a tree there exists an arrangement π' satisfying $\Omega(\pi', A) \leq \Omega(\pi, A)$ with the same or less amount crossings (Shiloach, 1979).

We define the ratio

$$\Gamma = \Omega_{real}/\Omega_{mla}$$

where Ω_{real} and Ω_{mla} are, respectively, the average value of Ω for the Romanian collection of sentences and that of the corresponding m.l.a.'s. Γ is a growing function of n, the sentence length (Fig. 8.3). Therefore, the shorter the sentence, the higher the validity of the Euclidean distance minimization principle.

A lot of research has been devoted to find the universal patterns of languages spoken on Earth. Probably, the most popular of them concerns word order. Greenberg (Greenberg, 1968) classified languages according to the way they ordered the subject (S), the verb (V) and the object (O). There are six possible combinations of these elements: OSV, OVS, VOS, VSO, SOV and SVO.

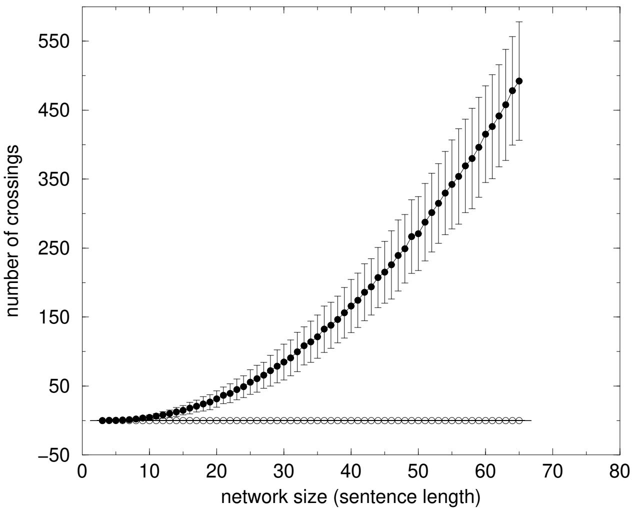

Figure 8.2: The average value of C, the number of arc crossings as a function of the network size, n, for random trees and their corresponding m.l.a. C grows faster than for random trees (black circles) whereas it remains very small for minimum linear arrangements (white circles) where $C < 0.022$ for ($n \leq 65$). Error bars length for m.l.a.'s are smaller than the mean values and thus not shown. Averages over 500 replicas are shown.

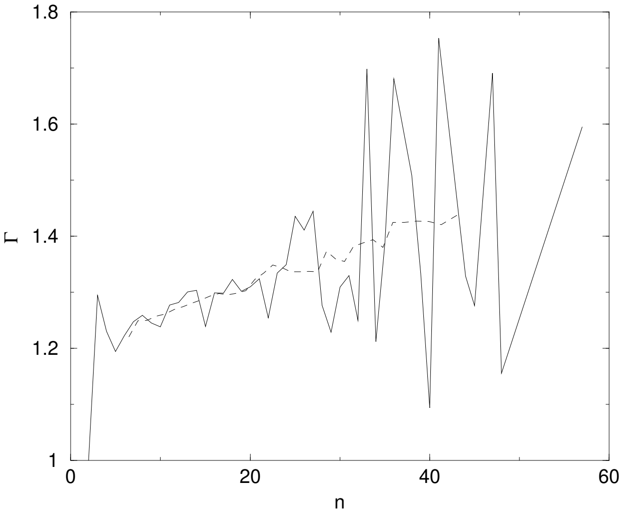

Figure 8.3: The optimization ratio Γ versus sentence length (solid line). Running averages show a tendency of Γ to grow with n (dashed line).

Surprisingly, only the last three are not rare in languages known (Greenberg, 1968; Dryer, 1989). SVO is a special configuration. Although SOV is twice more abundant than SVO (Dryer, 1989), different parsing models predict that SOV languages are harder to process (Pritchet, 1992; Babyonyshev and Gibson, 1999). Additionally, SVO is among the alternative word orders when VOS, VSO and SOV are the dominant orders (Steele, 1978). Furthermore, Bickerton put a step forward and stated there is a strong universal preference for SVO, which was evident in pidgin and creole languages studies (Bickerton, 1981). Why are some configurations preferred? We will see the m.l.a. can make predictions concerning S-V-O possible orderings under certain conditions. Notice that the basic word order chosen for a language by the generative tradition is the one providing the most economical description of sentences in that language (Derbyshire, 1977). Such basic order is not necessarily the most frequently used in real sentences, where the Euclidean distance minimization principle operates. We will hereafter implicitly assume we are referring to the most frequent word order.

Knowing that the network structure of the triple is $A = (S, V), (O, V)$ we may write Ω_{xyz} for $\Omega(\pi, A)$ with $\pi(x) = 1$, $\pi(y) = 2$, $\pi(x) = 3$ where $x, y, z \in \{S, V, O\}$ and $x \neq y \neq z$. An m.l.a. can very easily explain why SVO is preferred when S, V and O are formed by just one word each. We have that

$$\Omega_x = \begin{cases} 2 & if\ x \in \{SVO, OVS\} \\ 3 & otherwise. \end{cases} \tag{8.3}$$

We may consider the general case where the structure of S,V and O are the trees T_S, T_V and T_O, respectively. The whole sentence is formed by linking the head word of T_S with the head word of T_V and the head word of T_O with the head word of T_V. We define Ω_S, Ω_V and Ω_O as the sum of the distance between linked vertices in T_S, T_V and T_O, respectively. Assuming Ω_S, Ω_V and Ω_O do not depend on the type of S-V-O arrangement, we may write

$$\delta_x = \Omega_x - \Omega_S - \Omega_V - \Omega_O \tag{8.4}$$

where x is whatever S-V-O arrangement. The condition $\Omega_{SVO} < \Omega_x$ becomes $\delta_{SVO} < \delta_x$ for $x \neq SVO$. Assuming edges do not cross, we define L_x and R_x as the number of vertices on the left and on the right of the head of T_x, where $x \in \{S, V, O\}$. The number of vertices of T_x is $L_x + R_x + 1$. We have

$$\begin{aligned} \delta_{SVO} &= R_S + L_V + R_V + L_O + 2 \\ \delta_{SOV} &= 2L_V + 2R_O + L_O + R_S + 3 \\ \delta_{VSO} &= 2R_V + 2L_S + L_O + R_S + 3 \\ \delta_{OSV} &= 2R_S + 2L_V + R_O + L_S + 3 \\ \delta_{VOS} &= 2R_V + 2L_O + R_0 + L_S + 3 \\ \delta_{OVS} &= R_O + L_V + R_V + L_S + 2. \end{aligned} \tag{8.5}$$

We define $\Delta_x = R_x - L_x$ with $x \in \{S, V, O\}$. The condition $\delta_{SVO} < \delta_{SOV}$ leads to

$$\Delta_V < 2R_O + 1. \tag{8.6}$$

The condition $\delta_{SVO} < \delta_{VSO}$ leads to

$$-\Delta_V < 2L_S + 1. \tag{8.7}$$

The condition $\delta_{SVO} < \delta_{OSV}$ leads to

$$\Delta_V < \Delta_O + L_S + 1. \tag{8.8}$$

The condition $\delta_{SVO} < \delta_{VOS}$ leads to

$$\Delta_S < \Delta_V + R_O + 1. \tag{8.9}$$

The condition $\delta_{SVO} < \delta_{OVS}$ leads to

$$\Delta_S < \Delta_V. \tag{8.10}$$

The previous inequalities have interesting properties. Eq. 8.6 is trivially satisfied when $\Delta_V < 0$ since $R_O \geq 0$. Eq. 8.7 is trivially satisfied when $\Delta_V > 0$ since $L_S \geq 0$. Thereafter, if Eq. 8.6 is satisfied with $\Delta_V \geq 0$, then Eq. 8.7 is trivially satisfied. Inversely, if Eq. 8.7 is satisfied with $\Delta_V \leq 0$, then Eq. 8.6 is trivially satisfied. In other words, there are conditions where Bickerton's SVO universal preference follows trivially.

In order to investigate how much better is SVO versus the remaining orders, all distinct possible configurations of $(L_S, R_S, L_V, R_V, L_O, R_O)$ obeying

$$L_S + R_S + L_V + R_V + L_O + R_O + 3 = n \tag{8.11}$$
$$L_S, R_S, L_V, R_V, L_O, R_O \geq 0 \tag{8.12}$$

(where n is the length of the sentence) were generated for different values of n. p, the proportion of configurations where SVO was better than the remaining arrangements according to Eq. 8.6,8.7,8.8,8.9,8.10 was calculated (Fig. 8.4). If all different configuration of $(L_S, R_S, L_V, R_V, L_O, R_O)$ have the same frequency there are only three types of equations with respect to p, i.e. class I with Eq. 8.6,8.7, class II with Eq. 8.8,8.9 and class III with Eq. 8.10. In other words, we have class I for SOV and VSO, class II for OSV and VOS and class III for OVS. Thus, the complexity of the SVO preference problem has been reduced from five opponents to just three. As expected, we have $p = 1$ for $n = 3$ (as shown above) and $p \geq 0.66$ for $n \geq 3$ (Fig. 8.4) except for class III. Notice that we are assuming that all configurations of $(L_S, R_S, L_V, R_V, L_O, R_O)$ satisfying Eq. 8.12 have the same probability. Thereafter, this does not contradict that SOV is twice more abundant than SVO in real languages. Real languages clearly have biased the probabilities of $(L_S, R_S, L_V, R_V, L_O, R_O)$ configurations and differ in the amount of symmetry breaking (Jenkins, 2000). The only case where our prediction could be fully observed is in pidgin-creole languages. Pidgin speakers have their own mother tongue (different than the pidgin), which may introduce some bias in the word orders selected. The repertoire of word orders used by pidgin languages is heterogeneous (Arends, Muysken, and Smith, 1995). In contrast, SVO ubiquity is overwhelming in creole languages (Arends, Muysken, and Smith, 1995), suggesting that the initial stage for creoles are configurations having (almost) the same probability. This is consistent with the fact that creoles have native speakers and pidgins do not (pidgin speakers are biased by their mother tongue).

The Euclidean distance minimization principle should be more clear at the very beginning of new languages and our findings strongly suggest that SVO should have been the word order chosen when human language appeared on Earth. SOV has been proposed as earliest basic word order (Newmeyer, 2000). Assuming that OVS is not a regular candidate word order we hypothesize and will informally show *ad absurdum* that a proto-word-order other than SVO is less likely. OVS is not a regular candidate for a proto-word-order. Besides the fact there are a very few OSV languages, mostly in Amazonia (Derbyshire, 1977), OVS is a worst case situation for discourse organization since topics tend to appear early in the sentence and subjects are highly topical (Li, 1976; Givón, 1979). If a proto-word-order other than SVO (and OVS) is assumed then there are two possibilities

1. Euclidean distance minimization is not the principle governing the distance between syntactically dependent words. This would have two consequences:

   (a) Dependency trees would usually have crossings (if $n > 3$), which is a strictly universal regularity in present-day languages. As far as we know, there is no S-V-O arrangement where crossings are usual, but this is not an ultimate answer. There are two better reasons. Projectivity facilitates the analysis and synthesis of sentences (Melčuck, To appear). Thereafter, violating it is not encouraged. Nonetheless, there is a way of not minimizing the length of sentences and keep sentence structure without crossings: using sentences of length $1 \leq n \leq 3$, since all their dependency trees are trivially non-crossing. In that case, language does not qualify for the earliest language on Earth, since it can not make use discrete infinity. This kind of short sentence speech is found in children and trained primates (Johnson, Davis, and Macken, 1999; Gardner and Gardner, 1994).

   (b) Why the early hominids were endowed with a brain that allowed them to overcome the cost of not minimizing the distance between syntactically linked words must be explained (Gibson, 2000; Hawkins, 1994). Have nowadays brains decreased this capacity to link distant words with respect to hominids? It seems unlikely.

2. $(L_S, R_S, L_V, R_V, L_O, R_O)$ configurations are not equally likely. There are only two possible explanations for the bias:

   (a) The existence of an established language providing with the biases. This contradicts we are modeling the earliest language with subject, verb and object on Earth.

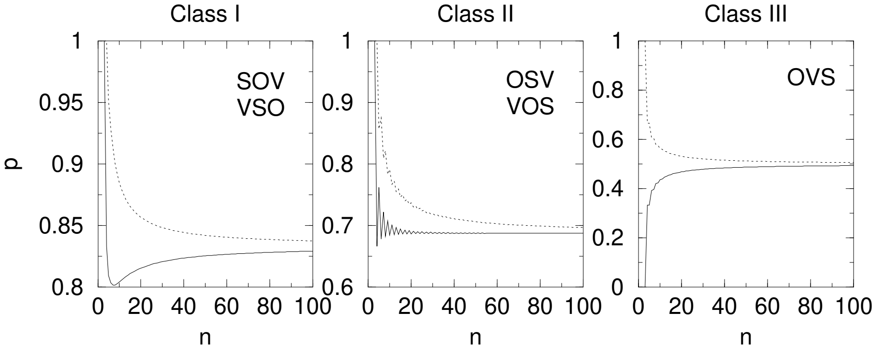

Figure 8.4: p, the proportion of configurations were SVO is more economical than the remaining configurations versus n. There are only three types of behavior: class I for SOV and VSO (left), class II for OSV and VOS (center) and class III for OVS. $p \geq 0.66$ for class I and II. Solid and dotted lines indicate, respectively, strict and non-strict satisfaction of the SVO superiority inequalities.

(b) Other forces favoring configurations other than SVO. This possibility should explain why such forces would be overcome in pidgin-creole languages, where SVO is universally preferred. It follows from the negative correlation between both degree of optimization and sentence length (Fig. 8.3) and the negative correlation between SVO superiority and sentence length (Fig. 8.4) that preference for SVO would have been maximal for the shortest sentences. If the first language spoken on Earth increased mean sentence length over time as children do over age (Reich, 1986), SVO preference would have been maximal at the very beginning.

The possibility that projectivity is violated because Euclidean distance minimization was not a principle at the emergence of language can not be denied. Nonetheless, if such a principle is assumed, the positive correlation between Γ and n suggests that factors preventing Ω from achieving a minimal value increase its effect as n grows. Grammatical rules could be one of them. Languages fixate precedence orders that the m.l.a can not supersede. For instance, Romanian rules that adjectives usually follow the noun (whereas English rules that adjectives must preceded its noun). Our results suggest the higher the length of the sentence, the higher the amount of precedence rules that must be obeyed.

Precedence rules may have an m.l.a motivation, but it is interesting to understand how precedence rules may result from no kind of optimization, as the following *in silico* experiments shows. A consensus compositional (Kirby, 2000) or recursive grammar (Kirby, 2002a) may emerge through computer simulations in a population of interacting agents without the need of Darwinian selection. The S-V-O order that emerges differs from run to run. The emerging of a S-V-O order is the result of a symmetry breaking process among all possible configurations (Jenkins, 2000). The experiment is a particular case of path dependent process (Arthur, 1994) where an arrangement is chosen not according to how optimal it is but according to how often it has previously been used. The m.l.a. must compete against not only precedence rules, but also against the agreement reached by a population. If forces other than m.l.a are very high and impose a suboptimal arrangement, the m.l.a. may still survive under the form of compensations for minimizing the cost of the arrangement. This is the hypothesis defended and successfully verified by M. Ueno on a SOV language. She hypothesizes, all languages are designed to be equally easy to process, regardless of the basic word order. She shows that Japanese (SOV), a suboptimal language with regard to SOV (assuming all configurations are equally likely), has a significantly greater abundance of one place predicates than English (SVO). Consistently, if predicate arguments are assigned to the verbal phrase V, a greater abundance of one place predicates implies a smaller L_V, leading to a decrease in $|\Omega_{SVO} - \Omega_{SVO}|$ and higher chance that Eq. 8.6 is not satisfied.

Children have more limited resources than adult speakers for language (Newport, 1990). The optimization principle discussed here should be specially evident in children utterances. Consistently, children sentences are the sentence set giving the greatest support for M. Ueno's hypothesis of compensations when a suboptimal S-V-O arrangement is chosen (Ueno and Polinsky, 2002). Moreover, the positive correlation between Γ and n predicts that children sentences should be the ones where optimization and compensations hypothesis should be observed, since mean sentence length is a growing function of age (Reich, 1986).

To sum up, the Euclidean distance minimization principle has to compete against multiple factors (not aiming to be exhaustive):

- The precedence rules a language has determined.
- The influence of the speaker's mother tongue in creole languages.
- Consensus. A certain word order might be optimal than other but not accepted by the majority of the population (Kirby, 2000; Kirby, 2002a).
- Psychological preferences for certain word orders. Topical information tend to generally occur early in a sentence (Li, 1976; Givón, 1979).

A m.l.a. has to compete against other forces so there are some conditions where it should be clearly observed:

- Short sentences (Fig. 8.3).
- Children's speech (Ueno and Polinsky, 2002).
- Creole languages. Pidgins do not have native speakers but creoles does (Arends, Muysken, and Smith, 1995). The overhelming use of SVO by creoles suggest that first infant creole speakers start from an unbiased initial state.

- The first language spoken on Earth, since it should be less influenced by biases such as word precedence order (recall the *ad absurdum* informal proof above).

The distance between linked words an ideal language subject to the Euclidean distance minimization principle can be calculated. Assuming the the m.l.a. is not distorted by other factors, we can predict $P(d)$, the probability that two linked words are at distance d. Using the maximum entropy principle (Kapur, 1989a) for obtaining $P(d)$ when the arc mean length $< d >$ is minimized, as Fig. 8.1 suggests, we get (Appendix F)

$$p_d = a(n-d)e^{-\beta d} \tag{8.13}$$

where β is a parameter and

$$a = \left(\sum_{d=1}^{n-1} (n-d)e^{-\beta d} \right)^{-1}.$$

β is a parameter satisfying

$$< d > = \sum_{d=1}^{n-1} d(n-d)e^{-\beta d}.$$

For large n (see Appendix F) we have

$$\beta \approx \frac{n- < d > \pm(< d >^2 -10n < d > -n^2)^{1/2}}{2 < d > n} \tag{8.14}$$

and n and $< d >$ are the only parameters.

If real sentences obeyed a full m.l.a. in full, a straight line in linear-log scale with the predicted exponent would be expected. While Eq. 8.13 is close to the real values for short distances, it can not directly explain the exponential trend with different slope for long distances (Fig. 8.5). The slower decrease in $P(d)$ of real sentences suggests the presence of factors such as precedence rules preventing $P(d)$ to decrease as fast as a pure m.l.a would dictate. We may use Eq. 8.1 as a null hypothesis for the expected distribution when distances for links are chosen regardless of the distance. The null hypothesis clearly differs from the real value of $P(d)$. Since creoles are mostly SVO, our results suggest that the agreement of creoles (at the earliest stages) languages with the functional $P(d)$ above should be higher than in regular languages.

8.4 Discussion

In short, inspired by Ueno's proposal (Ueno and Polinsky, 2002) we hypothesize that all human languages are strongly subject to an Euclidean distance minimization principle. Such a principle makes the following successful predictions:

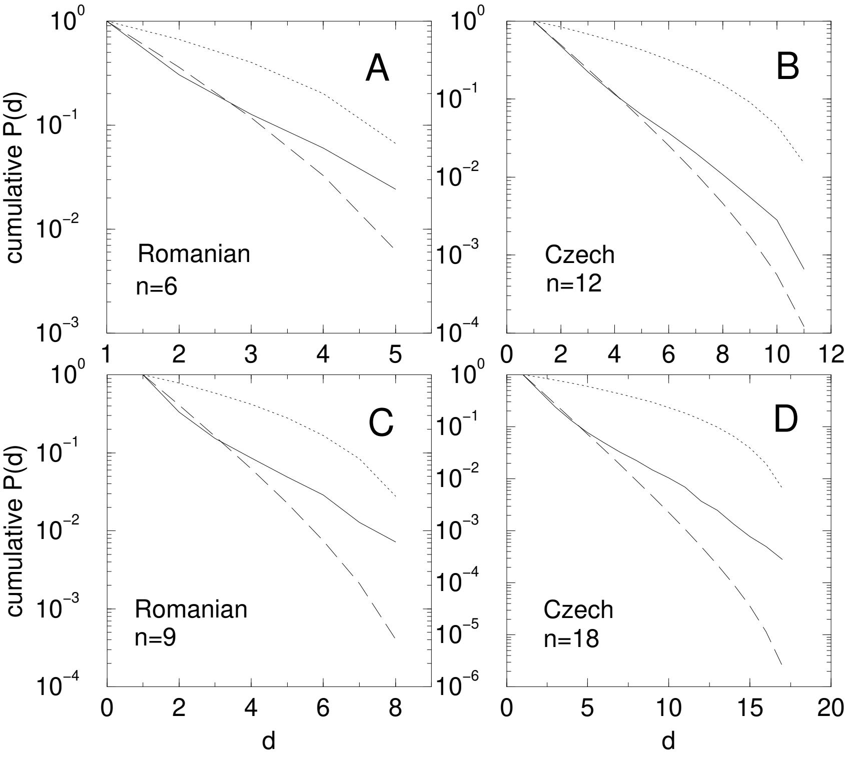

Figure 8.5: The cumulative $P(d)$, where $P(d)$ is the probability an arc links words at distance d. Real values (solid lines) can be compared to that of the null hypothesis (dotted lines) and the maxent exponential model (dashed lines). A. Romanian sentences having the typical length $L^* = 6$. B. Czech sentences having the typical length $L^* = 12$. C. Romanian sentences having the mean length $< L >\approx 9$ D. Czech sentences having the mean length $< L >\approx 18$. Real $P(d)$ clearly differs from the null hypothesis and approaches a straight line in linear-log scale, agreeing with the exponential prediction derived in this chapter for short distances and changing the slope but keeping the exponential trend for long distances.

- The projectivity constraint.

- An exponential distribution in the distance between linked pairs

- SVO has a wider efficiency (i.e. $p > 0.66$) than the remaining word orders. The need of SVO for using alternative word orders is smaller than the remaining word orders. The remaining word orders should make use of alternative word orders to overcome its limitations. Consistently, SVO has, in general, no common alternative orders whereas VOS, VSO and SOV do (Steele, 1978). Not surprisingly, VOS, VSO and SOV have always SVO as their alternative word order.

- The presence of compensations when the Euclidean distance minimization principle is overcome by other forces. Compensations should be stronger for class I order where $p \geq 0.8$ than for class II orders where $p \geq 0.66$.

- The way sentences must be arranged for achieving low values of Ω whenever SVO is not chosen (see Eqs. 8.6,8.7,8.8,8.9,8.10).

- SVO is a privileged candidate for the order used by the first (syntactic) language spoken on Earth.

Our work puts a step forward for understanding how optimization shapes the structure of sentences from different points of view. Distance minimization does not simply make current models more real. It explains many universal features. Our work limits the scope of innateness for explaining language universals (Hawkins, 1992). Languages spoken on Earth exhibit them because having limited resources is universal among human speakers.

Chapter 9

Network distance minimization

9.1 Introduction

Many essential features displayed by complex systems, such as memory, stability and homeostasis emerge from their underlying network structure (Strogatz, 2001; Kauffman, 1993). Different networks exhibit different features at different levels but most complex networks are extremely sparse and exhibit the so-called small-world phenomenon (Watts and Strogatz, 1998). An inverse measure of sparseness, the so-called network density, is defined as

$$\rho = \frac{\langle k \rangle}{n-1} \tag{9.1}$$

where n is the number of vertices of the network and $\langle k \rangle$ is its average degree. For real networks we have $\rho \in [10^{-5}, 10^{-1}]$ [1].

It has been shown that a wide range of real networks can be described by a degree distribution $P(k) \sim k^{-\gamma}\phi(k/\xi)$ where $\phi(k/\xi)$ introduces a cut-off at some characteristic scale ξ. Three main classes can be defined (Amaral et al., 2000). (a) When ξ is very small, $P(k) \sim \phi(k/\xi)$ and thus the link distribution is single-scaled. Typically, this would correspond to exponential or Gaussian distributions; (b) as ξ grows, a power law with a sharp cut-off is obtained; (c) for large ξ, scale-free nets are observed. The last two cases have been shown to be widespread and their topological properties have immediate consequences for network robustness and fragility (Barabási and Albert, 2002). The three previous scenarios are observed in: (a) power grid systems and neural networks (Amaral et al., 2000), (b) protein interaction maps (Jeong et al., 2001), metabolic pathways (Jeong et al., 2000) and electronic circuits (Ferrer i Cancho, Janssen, and Solé, 2001) and (c) Internet topology (Jeong et al., 2000; Caldarelli,

[1] Statistics performed on Table I in Ref. (Barabási and Albert, 2002)

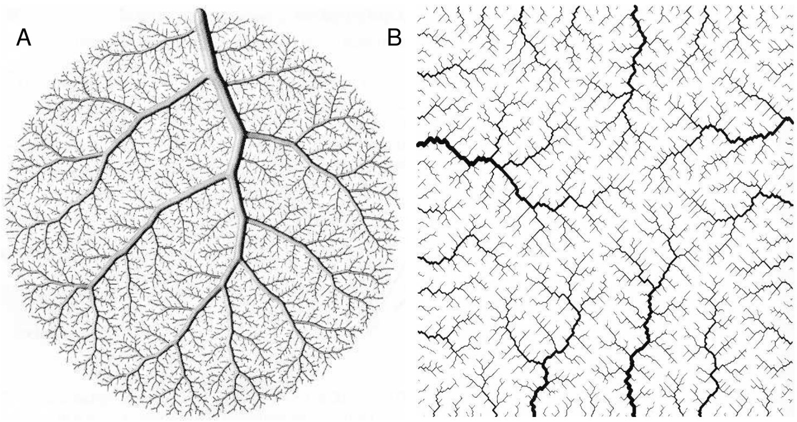

Figure 9.1: Optimal transport networks in biology (A) and geomorphology (B). A. An optimal tree structure that has been obtained for a vascular system on a two dimensional perfusion area (Brown and West, 2000). B. An optimal river basin network (also displaying tree structure) that has been generated by minimizing energy expenditure (Rodriguez-Iturbe and Rinaldo, 1997).

Marchetti, and Pietronero, 2000), scientific collaborations (Newman, 2001) and linguistic networks (Chapter 6).

9.2 Network optimization

Scale-free nets are particularly relevant due to their extremely high homeostasis against random perturbations and fragility against removal of highly connected nodes(Albert, Jeong, and Barabási, 2000). These observations have important consequences, from evolution to therapy (Jeong et al., 2001). One possible explanation for the origin of the observed distributions would be the presence of some (decentralized) optimization process.

Network optimization is actually known to play a leading role in explaining allometric scaling in biology (West, Brown, and Enquist, 1997; Brown and West, 2000; Banavar, Maritan, and Rinaldo, 1999) and has been shown to be a driving force in shaping neural wiring at different scales (Cherniak, 1995; Mitchinson, 1991) (see also (Bornholdt and Sneppen, 2000)). In a related context, local and/or global optimization has been also shown to provide remarkable results within the context of channel networks (Rodriguez-Iturbe and Rinaldo, 1997). By using optimality criteria linking energy dissipation and runoff production, the fractal properties in the model channel nets were essentially indistinguishable from those observed in nature. Fig. 9.1 displays different optimal transportation networks.

Several mechanisms of network evolution lead to scale-free structures within

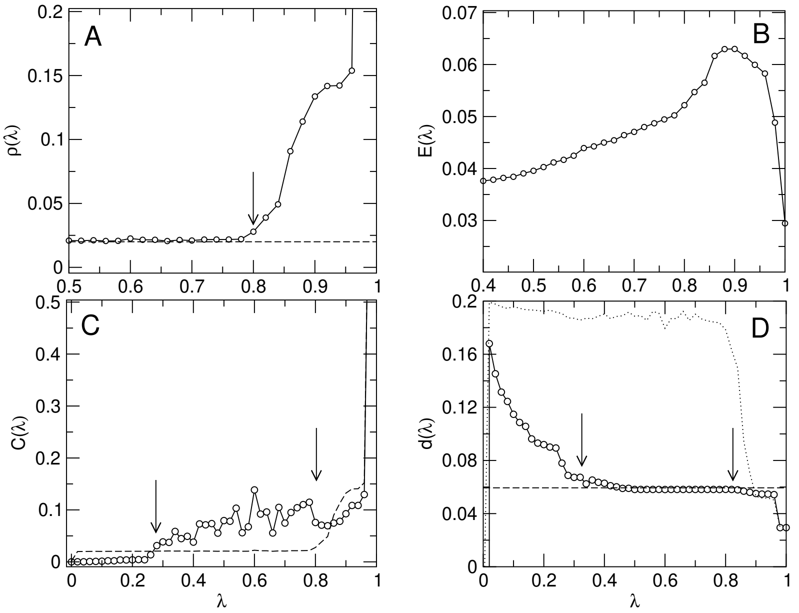

Figure 9.2: Density (A), energy (B), clustering coefficient (C) and distance (D) as a function of λ. Averages over 50 optimized networks with $n = 100$, $T = \binom{n}{2}$, $\nu = 2/\binom{n}{2}$ and $\rho(0) = 0.2$ are shown. A: the optimal network becomes a complete graph for λ close to 1. The density of an ideal star network, $\rho_{star} = 2/n = 0.02$ is shown as reference (dashed line). The clustering coefficient of a Poissonian network $C_{random} = \langle k \rangle/(n-1)$ is shown as reference in C. Notice that $C_{random} \approx \rho$. The normalized distance of a star network is (see Appendix E), $d_{star} = 6(n-1)/(n(n+1)) = 0.058$ (dashed line) and that of a Poissonian network, $d_{random} = \log n / \log \langle k \rangle$ (dotted line) are shown for reference in D.

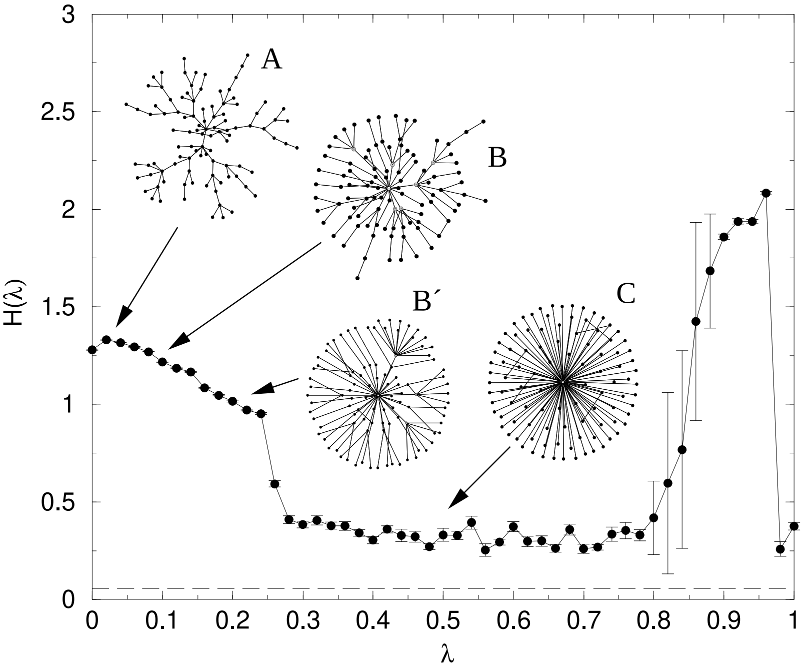

Figure 9.3: Average (over 50 replicas) degree entropy as a function of λ with $n = 100$, $T = \binom{n}{2}$, $\nu = 2/\binom{n}{2}$ and $\rho(0) = 0.2$. Optimal networks for selected values of λ are plotted. The entropy of a star network, $H_{star} = \log n - [(n-1)/n] \log(n-1) = 0.056$ is provided as reference (dashed line). A: an exponential-like network with $\lambda = 0.01$. B: A scale-free network with $\lambda = 0.08$. Hubs involving multiple connections and a dominance of nodes with one connection can be seen. C: a star network with $\lambda = 0.5$. B': a intermediate graph between B and C in which many hubs can be identified.

the context of complex networks in which the only relevant elements are vertices and connections (Barabási and Albert, 1999). Optimization has not been found to be one of them (Barabási and Albert, 2002). In this context, it was shown that (Metropolis-based) minimization of both vertex-vertex distance and link length (*i.e.* Euclidean distance between vertices)(Mathias and Gopal, 2001) can lead to the small-world phenomenon and hub formation. This view takes into account Euclidean distance between vertices. Here we show how minimizing both vertex-vertex distance and the number of links leads (under certain conditions) to the different types of network topologies depending on the weight given to each constraint. These two constraints include two relevant aspects of network performance: the cost of physical links between units and communication speed among them.

9.3 The optimization algorithm

For the sake of simplicity, we take an undirected graph having a fixed number of nodes n and links defined by a binary adjacency matrix $A = \{a_{ij}\}$, $1 \leq i, j \leq n$. Given a pair of vertices i and j, $a_{ij} = 1$ if they are linked ($a_{ij} = 0$ otherwise) and D_{ij} is the minimum distance between them. At time $t = 0$, we have a randomly wired graph (i.e. a Poisson degree distribution) in which two given nodes are connected with some probability p. The energy function of our optimization algorithm is defined as

$$\Omega(\lambda) = \lambda d + (1 - \lambda)\rho$$

where $0 \leq \lambda, d, \rho \leq 1$. λ is a parameter controlling the linear combination of d and ρ. The normalized number of links (i.e. the link density), ρ is defined in terms of a_{ij} as

$$\rho = \frac{1}{\binom{n}{2}} \sum_{i<j} a_{ij}$$

and it is equivalent to Eq. 9.1. The normalized vertex-vertex distance, d, is defined as $d = D/D^{linear}$ being

$$D = \frac{1}{\binom{n}{2}} \sum_{i<j} D_{ij}$$

the average minimum vertex-vertex distance and $D^{linear} = (n + 1)/3$ the maximum value of D that can be achieved by a connected network, that is, that of a linear graph (see Appendix E).

We define a linear graph as a graph having 2 vertices with degree 1 and $n-2$ vertices with degree 2 [2]. A graph whose adjacency matrix satisfies

$$a_{ij} = \begin{cases} 1 & \text{if } |i - j| = 1 \\ 0 & \text{otherwise} \end{cases} \tag{9.2}$$

is a linear graph. Such a graph has the maximum average vertex-vertex distance that can be achieved by a connected graph of order n (see Appendix E).

The minimization of $\Omega(\lambda)$ involves the simultaneous minimization of distance and number of links (which is associated to cost). In other words, Ω is a combination of two principles, a network distance principle and a density minimization principle. Notice that minimizing $\Omega(\lambda)$ implies connectedness (*i.e.* finite vertex-vertex distance) except for $\lambda = 0$, where it will be explicitly enforced.

The minimization algorithm proceeds as follows. At time $t = 0$, the network is set up with a density $\rho(0)$ following a Poissonian distribution of degrees (connectedness is enforced). At time $t > 0$, the graph is modified by randomly changing the state of some pairs of vertices. Specifically, with probability ν, each a_{ij} can switch from 0 to 1 or from 1 to 0. The new adjacency matrix is

[2] It can be easily shown through induction on n that such a graph is connected and has no cycles.

accepted if $\Omega(\lambda, t+1) < \Omega(\lambda, t)$. Otherwise, we try again with a different set of changes. The algorithm stops when the modifications on $A(t)$ are not accepted T times in a row. The minimization algorithm is a simulated annealing at zero temperature.

The basic scheme of the minimization algorithm is the same as in Chapter 3. Here, $\mathbf{A}$ is set up with a fixed density $\rho(0)$ of ones at the beginning of the procedure.

Hereafter, $n = 100$ [3], $T = \binom{n}{2}$, $\nu = 2/\binom{n}{2}$ and $\rho(0) = 0.2$.

We define the degree entropy on a certain value of λ as

$$H(\{p_k\}) = -\sum_{k=1}^{n-1} p_k \log p_k$$

where p_k is the frequency of vertices having degree k and $\sum_{k=1}^{n-1} p_k = 1$. This type of informational entropy will be used in our characterization of the different phases [4].

Some of the basic average properties displayed by the optimized nets are shown against λ in Fig. 9.2. These plots, together with the degree entropy in Fig. 9.3 suggest that four phases are present, separated by three sharp transitions at $\lambda_1^* \approx 0.25$, $\lambda_2^* \approx 0.80$ and $\lambda_3^* \approx 0.95$ (see arrows in Fig. 9.2). The second one separates sparse nets from dense nets and fluctuations in $H(\lambda_3^*)$ are specially high. $\rho(\lambda), C(\lambda) \approx 1$ for $\lambda > \lambda_3^* \approx 0.95$. For $\lambda = 0$ and $\lambda = 1$ a Poissonian and a complete ($\rho(\lambda) = 1$) network are predicted, respectively.

9.4 Optimal degree distributions

When taking a more careful look at the sparse domain $(0, \lambda_2^*)$, three non-trivial types of networks are obtained as λ grows:

a. Exponential networks, i. e. $P_k \sim e^{-k/\xi}$.

b. Truncated scale-free networks, i. e. $P_k \sim k^{-\gamma} e^{-k/\xi}$ with $\gamma = 3.0$ and $\xi \approx 20$ (for $n = 100$).

c. Star network phase ($\lambda_1^* < \lambda < \lambda_2^*$) *i.e.* a central vertex to which the rest of the vertices are connected to (no other connections are possible). Here,

$$p_k = \frac{n-1}{n}\delta_{k,1} + \frac{1}{n}\delta_{k,n-1} \tag{9.3}$$

[3] Higher values of n were very time consuming. The critical part of the algorithm is the calculation of d which has cost $\Theta(n\rho\binom{n}{2})$, that is, $\Omega(n^2)$ and $O(n^3)$. Faster calculation implies performing an estimation of d on a random subset of vertices or 1st and 2nd neighbors (Newman, Strogatz, and Watts, 2001) that happened to be misleading.

[4] Entropy measures of this type have been used in characterizing optimal channel networks and other models of complex systems (see (Solé and Miramontes, 1995)) although they are typically averaged over time.

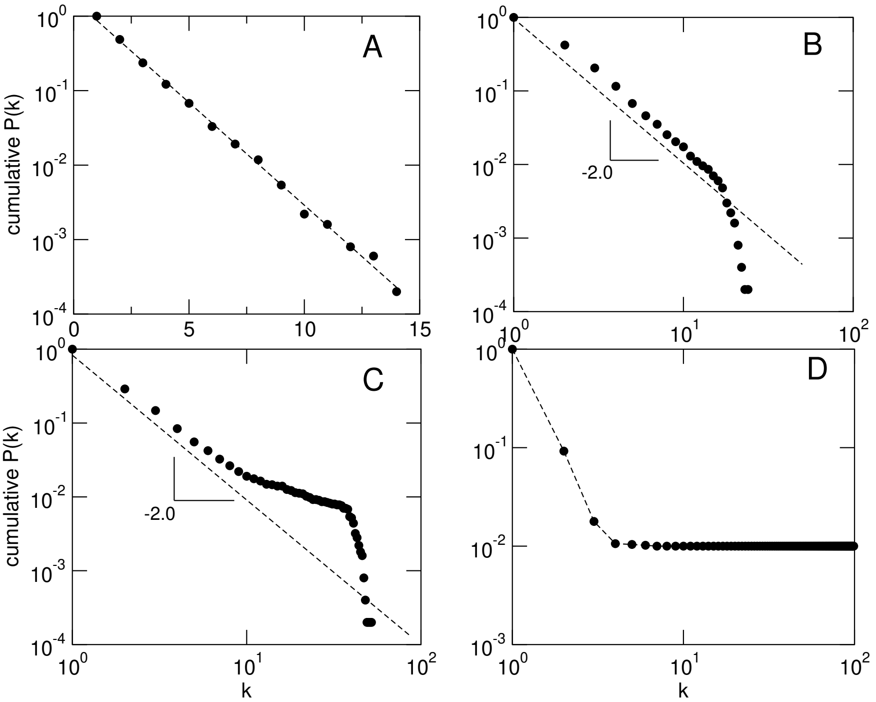

Figure 9.4: Selected cumulative degree distributions of networks obtained minimizing $\Omega(\lambda)$. Every distribution is an average over 50 optimized networks with $n = 100$, $T = \binom{n}{2}$, $\nu = 2/\binom{n}{2}$ and $\rho(0) = 0.2$. A: an exponential-like distribution for $\lambda = 0.01$. B: a power distribution with exponent $\gamma = 2.0$ for $\lambda = 0.08$ (with a sharp cutoff at $\xi \approx 20$). C: $\lambda = 0.20$. D: $\lambda = 0.50$ (almost an star graph).

> A star graph has the shortest vertex-vertex distance between vertices among all the graphs having a minimal amount of links (see Appendix E). Consistently, star graphs are obtained when λ is sufficiently large here or using genetic algorithms in a similar context (Nishikawa et al., 2002). Non-minimal densities can be compensated with a decrease in distance, so pure star networks are not generally obtained here.

The distributions of (a-c) types and that of a dense network are shown in Fig. 9.4. A detailed examination of the transition between degree distributions reveals that hub formation explains the emergence of (b) from (a), hub competition (b') precedes the emergence of a central vertex in (c). The emergence of dense graphs from (c) consists of a progressive increase in the average degree of non-central vertices and a sudden loss of the central vertex. The transition to the star net phase is sharp. Figure 9.3 shows $\langle H(\lambda) \rangle$ along with plots of the major types of networks. It can be seen that scale-free networks (b) are found close to λ_1^*. The cumulative exponent of such scale-free networks is two and thus $\gamma = 3.0$, the same that it would be expected for a random network generated with the Barabási-Albert model (Barabási and Albert, 1999).

Our scenario suggests that preferential attachment networks might emerge at the boundary between random attachment networks (a) and forced attachment (*i.e.* every vertex connected to a central vertex) networks (c) and points that optimization can explain the selection of preferential attachment strategies in real complex networks. In our study, exponential-like distributions appear when distance is minimized under high density pressure, in agreement with the study by Amaral and co-workers on classes of small-world networks (Amaral et al., 2000). This might be the case of the power grid and of neural networks (Amaral et al., 2000). If linking cost decreases sufficiently, cliquishness becomes an affordable strategy for reducing vertex-vertex distance. Consistently, graphs tend to a complete graph for high values of λ. The Watts model (Watts and Strogatz, 1998) is a non-trivial example of what cliquishness (i.e. high clustering) can do for smallwordness. High clustering favours small-worldness (as seen for $\lambda \geq \lambda_2^*$) but it is not the only mechanism (Dorogovtsev and Mendes, 2002).

We have seen the different optimal topologies depending on the value of λ. We are aimed at defining an absolute measure of optimality depending on λ we can use for ranking the different topologies. We define

$$\Gamma(\lambda) = \frac{1 - d(\lambda)}{\rho(\lambda)} \tag{9.4}$$

as such measure (Fig. 9.5 A). A sharp transition from sparse to dense networks is clearly observed for $\lambda \approx 0.8$. According to Fig. 9.5 A, the topology ranking becomes,

1. Star networks.
2. Scale-free networks.
3. Exponential networks.

4. Dense networks.

See the Appendix E for a summary of the basic features of the trivial topologies appearing in our study.

A simpler version of the previous scenario appears in the context of Poissonian graphs, where we define the optimality measure as S/ρ, where S is the number of vertices of the largest connected component and ρ is both the expected network density and the probability that a random pair of vertices are linked. Again, the maximum divides networks into disconnected networks and connected networks at high link expense (Fig. 9.5 B). $\rho \approx 0.8$ divides low cost strategies from high cost strategies as $\lambda = 0.8$ does in Fig. 9.5 A. Notice that the transition is smooth for the former and sharp for the latter. The Poissonian scenario shows the optimization principles that may guide networks in early stages to remain close to the connectedness transition. Once enough connectedness is achieved, networks may be guided by Eq. 9.4 or particular values of λ depending on the system.

9.5 Discussion

The network previous results and our conjecture concerning optimization in complex nets requires explaining why star graphs are not found in nature. Different constraints can be restricting the access of star graphs to real systems. Let us list some of them:

- Randomness. The evolution of the topology as λ grows suggests a transition from disorder (exponential degree distribution) to order (star degree distribution).

- Diversity. The number of different star graphs that can be formed with n vertices is n whereas it explodes for exponential and power distributions.

- Robustness. Removing the central hub leaves $n-1$ connected components, which is the worst case situation.

Whether or not optimization plays a key role in shaping the evolution of complex networks, both natural and artificial, is an important question. Different mechanisms have been suggested to explain the emergence of the striking features displayed by complex networks. Most mechanisms rely on preferential attachment-related rules, but other scenarios have also been suggested (Solé et al., 2002; Vázquez et al., 2003) in which external parameters have to be tuned. When dealing with biological networks, the interplay between emergent properties derived from network growth and selection pressures has to be taken into account. As an example, metabolic networks seem to result from an evolutionary history in which both preferential attachment and optimization are present. The topology displayed by metabolic networks is scale-free, and the underlying evolutionary history of these nets suggests that preferential attachment might have been involved (Fell and Wagner, 2000). Early in the evolution of life,

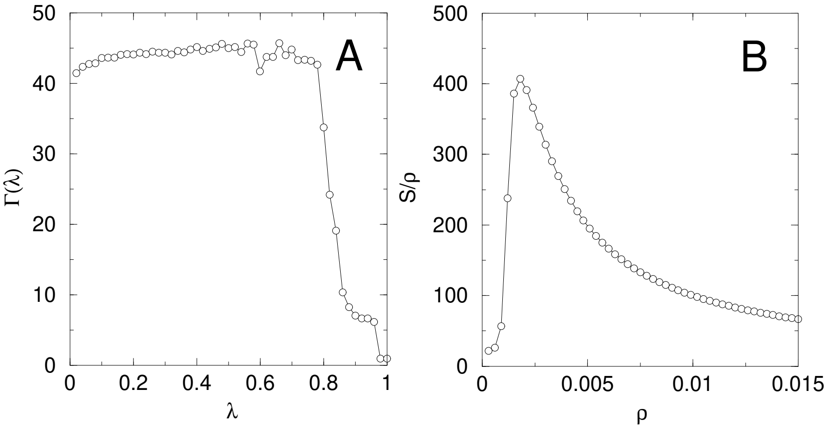

Figure 9.5: A. The function $\Gamma(\lambda) = (1 - d(\lambda))/\rho(\lambda)$ for the minimum energy configurations. B. The cost function S/ρ versus ρ for the Poissonian model.

metabolic nets grew by adding new metabolites, and the most connected are actually known to be the oldest ones. On the other hand, several studies have revealed that metabolic pathways have been optimized through evolution in a number of ways. This suggests that the resulting networks are the outcome of both contributions, plus some additional constraints imposed by the available components to the evolving network (Morowitz et al., 2000; Schuster, 2001). In this sense, selective pressures might work by tuning underlying rules of net construction. This view corresponds to Kauffman's suggestion that evolution would operate by taking advantage of some robust, generic mechanisms of structure formation (Kauffman, 1993).

The network morphospace resulting from the network distance minimization and the link density principle have implications for the global and sentence syntactic dependency networks. We have seen the statistical analysis of linguistic networks reveals the existence of at least two classes of networks. From the one hand, global syntactic dependency networks, resulting from collecting dependency links from a collection of sentence. Such networks obey a degree distribution

$$p_k \sim k^{-\gamma} \tag{9.5}$$

where $\gamma \approx 2.2$. Such networks exhibit disassortative mixing, that is, a tendency to avoid links between between vertices of a similar degree. From the other hand, an heterogeneous set of semantic networks obey Eq. 9.5 with $\gamma \approx 3$ but and exhibit assortative mixing, that is, a tendency to form links between vertices of a similar degree. We do not need to resort to the network distance and link density minimization principles for explaining the degree distributions of syntactic dependency networks. We have seen such distributions easily follow from (a) Zipf's law and a linear relationship between word frequency and word syntactic degree (Chapter 6) and (b) the distribution of the number of objects per signal, for sufficiently large k (Chapter 7). Since $\gamma \approx 2$ implies a more dense network with regard to $\gamma \approx 3$ (see Chapter 7), navigation must be faster in the former case, but as a side-effect of referential principles. As for semantic networks, the exponents found are close to the one obtained from network distance least effort and link density minimization principle in numerical calculations. Since network sizes in the numerical calculations are far from that of real semantic networks (the numerical exponent could depend on networks size) and optimization principles compete against preferential attachment principles (Barabási and Albert, 1999), the optimization principles presented in the current chapter can only be as possible explanations. Since fast communication pressures are at work in human language, the principles presented are interesting prospects for linguistic networks that have not been yet studied.

Network distance minimization can be regarded as an alternative hypothesis against Euclidean distance minimization for the degree distribution in the syntactic dependency structure of sentences, $P_{sentence}(k)$. We have seen that $P_{sentence}(k)$ in Chapter 6 is close to an exponential function. An exponential distribution has been obtained in the present chapter for small values of λ, that is when link density minimization is the most important force. Therefore, the present chapter suggests network distance minimization has little (if any) contribution to $P_{sentence}(k)$. Sentence structures minimize the Euclidean distance and not the network distance because sentence structure is mostly determined by discourse needs and the semantic relationships among the intervening words.

Chapter 10

Other views

The present chapter is devoted to show how the present thesis fits in three major fields, i.e. standard linguistics, the Chomskian tradition and general approaches to the origins of language. Devoting a special section to the Chomskian tradition is motivated by its influence on standard linguistics and even more important here, its influence on approaches to the origins of language. When researchers assume linguistic knowledge is innate (e.g. (Nowak and Krakauer, 1999)), a division between I-language and E-language (e.g. (Kirby, 2002a)), phrase structure as model of syntax (Kirby, 2002a), the existence of universal grammar (e.g. (Nowak, Komarova, and Niyogi, 2001)), that Zipf's law is meaningless (e.g. (Nowak, Plotkin, and Jansen, 2000)) or more widely, that there is no need to check models with real data because statistical patterns are deceiving, these are traces or of the Chomskian tradition.

10.1 Standard linguistics

Here we discuss two aspects of the present thesis that need to be clarified to the light of standard linguistics:

- *The nature of referentially useless words.*
  Standard linguistics distinguishes two major classes of words, i.e. content words and function words. A content words is usually defined as *a word to which an independent meaning can be assigned* whereas a function word is *a word that serves a grammatical function but has no identifiable meaning* [1]. Sometimes, it is just said the functions words have no meaning. One has to bear on mind that referential power and semantic content are not the same thing. The approach of the present thesis is that both kinds of words have semantic content. If semantic content is measured in terms of connections with objects of reference, then we may say that function

[1] Definitions borrowed from *www.hyperdictionary.com*. The English dictionary is based on WordNet 1.7.1 (by Princeton University)

words have the highest semantic content. One can clearly argue that function words are devoid of referential power, but one can not argue that function words are devoid of semantic content, because semantic emptiness and absence of referential power can not be distinguished, at least using $H(R|s)$, the entropy associated to the interpretation of a word s. We have seen in Chapter 4 that if a word s has μ meanings then $H(R|s) = \log \mu$. $\log \mu$ is therefore a measure of the decoding effort. Assuming that the frequency of use of a word, f, is correlated with μ, more precisely $f \sim \mu$, and that words are used according to their meaning, a referentially useless word can only come from

1. A word with no semantic content ($\mu = 0$), which contradicts the definition of a word since it will never be used ($f = 0$). Another contradiction comes from the fact that $H(R|s) = -\infty$ when $\mu = 0$, which contradicts that the effort for the hearer is maximal when a word has no associated objects of reference.
2. A word with $\mu > 0$. Therefore, the only way of obtaining referentially useless words is by means of too large values of μ. We know $H(R, s) \leq m$, where m is the amount of objects of reference.

To put it our words, here function words are still content words (have too meanings), but have too high $H(R|s_i)$.

- *Meaning versus objects of reference* The work presented has assumed in many places a specific understanding of the meaning of a word. Meaning is more than simple signal-object associations (Ravin and Leacock, 2000a). Our understanding of what words refers to is closer to Pulvermuller's where word forms are associated to different brain areas, which in turn are associated to different kinds of stimuli (motor, visual,...) (Pulvermuller, 1999; Pulvermuller, 2001). There are further reasons for realizing that our approach differs from the classical understanding of meanings that standard dictionaries mirror. If f is the frequency of a word and μ its number of meanings, we have assumed $f \sim \mu$ but studies where number of meanings is equated to number of entries in a dictionary lead to $f \sim \mu^{\upsilon}$ with $\upsilon = 1.76$ (Köhler, 1986) of $\upsilon \approx 2$ in early G. K. Zipf's studies (Manning and Schütze, 1999). Here we have chosen $\upsilon = 1$ for simplicity. Are we being inconsistent with real language? No. Dictionary entries are complex definitions involving the combination of more than one object of reference. Accordingly, frequency grows faster with the number of entries ($\upsilon \geq 1.7$ than with the number of objects ($\upsilon = 1$). Some of the findings here are independent of the definition of meaning assumed. For instance, the reader is proposed to change our abstract understanding of object of reference by its own definition of meaning and the main results here (Zipf's law in a sharp phase transition) will still be valid. To sum up, we can easily replace signals by words, but not objects by meanings. If two signals are linked to the same object they are not necessarily synonyms.

10.2 The Chomskian view

The Chomskian view makes some assumptions needing to be reconsidered:

- *Centralized versus emergent syntactic structures.*
The Chomskian tradition is based on phrase structure formalisms for describing the structure of sentences (Chomsky, 1957; Chomsky, 1995; Uriagereka, 1998; Jackendoff, 2002). Phrase structure formalisms consider that phrases are word groups that must be explicitly defined. The dependency grammar formalisms assume that phrases are epiphenomena of syntactic relationships between words ((Melčuk, 1988; Hudson, 1990)). Through the eyes of complexity, many patterns we observe in nature such as leopard stains (Murray, 1980) or shell patterns (Meinhardt, 1995) (Fig. 10.1) emerge from the interaction of individual components. There is no painter with a paintbrush. There is no central control of the boundaries of every component of the pattern. In contrast, phrase structure based formalisms control the boundaries of phrases explicitly and syntactic dependency based formalisms need to make a centralized decision about forbidding or allowing projectivity (Hudson, 1984; Melčuk, 1988).

- *The distance between formal descriptions and brain structure.*
Phrase structure grammar are models that are far from the structure of the brain. The brain itself is a network. Why the Chomskian tradition works on a model that is difficult to map on the brain? Syntactic networks are more suitable models since their mapping into brain structures is easier. Moreover, syntactic dependency networks share common patterns with the organization of the brain. The so-called small-world phenomenon is present in global syntactic dependency networks (Section 6) whereas neurons in the neocortex exhibit three degrees of separation (Braitenberg and Schuz, 1992).Networks match the brain structure and statistical patterns. What can be said about the similarities between rewriting rules and brain structure? Our view is close to the so-called 'cognitive linguistics' movement, which includes several specific theories: cognitive grammar (Langacker, 1987; Langacker, 1990), construction grammar (Goldberg, 1995; Kay and Fillmore, 1990) and word grammar (Hudson, 1990; Hudson, 1999a). These theories share the view that there is no boundary between the lexicon and the rules of grammar (Hudson, 1999b).

- *The importance of syntax.*
The Chomskian tradition defends the autonomy of syntax (Chomsky, 1957), which postulates that the crux of human language can be fully understood dissociated from other linguistic dimensions. Furthermore, the Chomskian tradition considers syntax is the most important aspect of human language faculty, but other views consider that symbolic reference is the most important aspect. Here we show that the two opposite views are correct because syntax and symbolic reference are two sides of the same coin (Chapter 7).

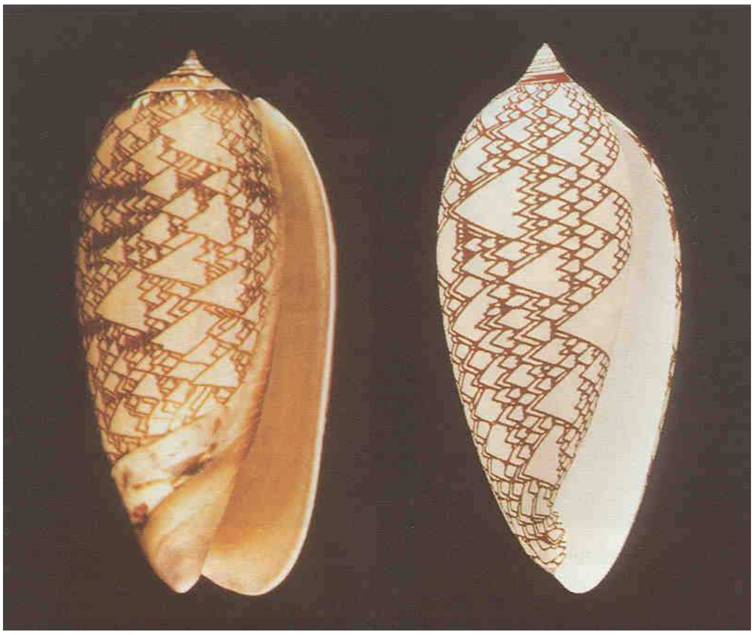

Figure 10.1: A photograph of *Oliva porphyria* and a model without central control. Reproduced from (Meinhardt, 1995).

- *Dualism versus integration.*
  The present thesis is a challenging investigation against the dualistic understanding of human language in the Chomskyan tradition:

  - The radical division between syntax and other levels (e.g. semantics), i.e. the autonomy of syntax. Such a division precludes integrative views where syntax and symbolic reference are two sides of the same coin, i.e. connectedness (Chapter 7). The chomskian takes a reductionist position with regard to syntax.

  - A radical distinction between competence and performance.

- *The risk of idealized models of language.*
  As for the latter, N. Chomsky hypothetically formulated the distinction between competence (the speaker's knowledge of language) and performance (the actual use of the language in concrete situations). Competence is an idealization considering memory limitations, distractions, shifts of attention and interest and errors, as irrelevant (Chomsky, 1965b). It has been shown here that different linguistic universals can be explained in terms of performance constraints. Human language performance is constrained by several factors, e.g. lung capacity imposes limits on the length of actual spoken sentences, whereas working memory imposes limits on the complexity of sentences is they are to be understandable (Hauser, Chomsky, and Fitch, 2002). All these limitations are considered to be outside of the language faculty (Hauser, Chomsky, and Fitch, 2002). It is clear competence provides abstraction which is in turn necessary for understanding the crux of human language. Since linguistic universals suggest, in Miller-Chomsky's words, *powerful and universal forces* (Miller and Chomsky, 1963), and competence is devoted to capture the essence of human language, one may wonder if competence is sufficient for explaining linguistic universals. That is not the case, at least for the universals studied here. We have seen human syntax and symbolic reference need a conflict between coding and decoding least effort (Chapter 7). Coding least effort is a word retrieval constraint inside our brain. Coding least effort is a performance constraint. In contrast, decoding least effort can be seen as a competence factor according to Chomsky's ontology. Projectivity is a consequence of an Euclidean distance minimization principle, which is in turn a consequence of our brain processing-memory limitations (Chapter 8). Euclidean distance minimization is a performance principle. Therefore, competence provides too level of abstraction for explaining linguistic universals. Most of XX century work on linguistics has consisted of providing researchers with ontologies upon which construct theories (Altmann, 1978). The competence-performance division belongs to such ontologies and needs to be reconsidered. There are two prospects, looking for intermediate levels of abstraction or probably better, integrative paths where it is admitted that different linguistic universals emerge from the the efficiency limitations of an ideal language faculty (Altmann, 1978).

We have seen projectivity is a side-effect of an Euclidean distance minimization principle. In other words, projectivity is an emergent feature of keeping distances between syntactically related words small. Considering projectivity a principle in the dependency grammar formalism is a mistake. That is why syntactic formalisms need to be careful when making the competence-performance division. The competence-performance division is necessary, but performance limitations shape the structure of sentences in non-trivial ways.

- *Zipf's law meaningfulness.*
  Theoretical approaches to language in the Chomskyan tradition are based on descriptions of linguistic phenomena, e.g. the structure of sentence and metadescriptions and principles concerning such descriptions. Simplicity and economy are examples of principles of descriptions and metadescriptions. In the 60s, G. Miller and N. Chomsky took previously reported evidence that intermittent silence models reproduced Zipf's law (with $\beta \approx 2$) (Miller, 1957; Mandelbrot, 1953) for arguing Zipf's law is a meaningless statistical regularity. N. Chomsky discouraged from then on research on statistical linguistic patterns [2]. If theoretical approaches to language avoid contact with real statistical patterns and psychological evidence (among others), the only possible way of filtering hypothesis are *a prioristic* principles such as simplicity. That is why the simplicity of intermittent silence made him the best model at that time. But the simplest explanation is not the best explanation. Simplicity is a desirable requirement but it is not qualifying from the point of view of the philosophy of science (Altmann, 1978). The sufficient condition for an explanation (obviously, besides being an explanation) is that its assumptions are valid. Models and even null hypothesis must be well grounded. As discussed earlier, this is not the case of intermittent silence models: words are chosen from a mental lexicon (not created from scratch every time), words are used according to their meaning, lexicon size is bounded by brain capacity. The work presented here must be understood as the end of a repeating reference to a badly-grounded explanation (Miller and Chomsky, 1963; Rapoport, 1982; Nowak, Plotkin, and Jansen, 2000; Nowak, 2000a; Nowak, 2000b; Wolfram, 2002; Suzuki, Tyack, and Buck, 2003).

- *The definition of complex communication.*
  The Chomskian tradition has its own definition of complex communication. If human language is a complex communication system, their approach is a differential method. What makes human language unique (and therefore complex) is the result of subtracting non-human species simple communication systems to human language. Therefore, recursion is what makes human language complex. Here, we have seen that Zipf's law (with $\beta \approx 2$) is the hallmark of complex communication, not only concerning the way signals and objects associate, but also the syntactic patterns that

[2] George Miller, personal communication

follow (Chapter 7). We therefore say that a communication system is complex when it is capable to handle the maximum tension between hearer and speaker needs. Our method is not differential and does not exclude the possibility that other species have complex communication systems if they exhibit Zipf's law. We have also seen that species have different ways of minimizing the decoding effort. Minimizing the total amount of meanings per word is a simple strategy. Minimizing the entropy associated to the interpretation of a words is a complex strategy (Chapter 4).

It is interesting to point that recursion is a weak indicator of complexity. Recursion has more to do with performance than with competence. The level of recursion has to do with brain capacity. Children have limited recursion skills, but their short sentences are still considered language. Therefore, the crucial trait is combinatorics (based on units with reference, see Section 1.3.1), that is, having the possibility of combining whatever pair of words. The length of the message is left to brain capacity and the necessary evolutionary gradual steps for increasing such capacity. Once combinatorics is achieved (i.e. connectedness) then the level of recursion is a quantitative metric. We have seen that connectedness is a side-effect of satisfying simple communicative needs. Therefore, our definition of complexity can not be ultimately defined in terms of connectedness but in terms of how the effort for the hearer and speaker is minimized. Combinatorics is a qualitative trait but recursion (assuming combinatorics is present) is a quantitative trait. Recursion is too specific. It is not focused on quality but on a quantity. Basing human language uniqueness on a high level of recursion seems an anthropocentric requirement for language designed to keep humans on top of the animal kingdom for the following reasons:

- Children have limited recursion.
- The communication systems of different species have not been deciphered or further understood (Section 1.3.1).
- Some species trained by humans exhibit limited recursion (Herman, Richards, and Wolz, 1984; Greenfield and Savage-Rumbaugh, 1991).

- *The nature of universality in language.*
  The Chomskian tradition hypothesizes the existence of a universal grammar. The universal grammar is devoted to be the backbone from which all existent languages on Earth stem. The universal grammar follows from another hypothesis, the existence of a language acquisition device, whose initial state would be the universal grammar (Uriagereka, 1998). Here we have hypothesized the existence of a single universality class for all languages on Earth. The arguments for the existence of the universal grammar are qualitative in nature, and mostly based on the poverty of stimulus paradox (Uriagereka, 1998). Our hypothesis of a single universality class is based on quantitative measures borrowed from statistical physics. We have shown that many potential syntactic universals arise

from from Zipf's law. Thereafter, our universality class relies on the two referential principles: coding and decoding least effort. Our approach finds essential similarities and explains them as a side effect of reference. Our approach to the essential similarities among languages on Earth is closed. In contrast, the universal grammar formulates hypothesis that are open. The validity of the universal grammar hypothesis and the language acquisition device is left to other disciplines ranging from learnability theory, typology and psycholinguistics. The Chomskian tradition is focused on the consequences of this hypothesis and not on the validity of their assumptions. But the nature of the universal grammar and the single universality class proposed here is innate. While there is an open debate about the origins of innateness in the context of universal grammar, our universality class is innate because of a matter of mathematical truth. Given a meaningful axiom, i.e. Zipf's law, syntactic dependency universals follow (Chapter 7). A language organ (Pinker, 1996) and constraint on what can be learned (Gold, 1967) are not necessary.

- *Actual word order versus a priori word order.*
  The Chomskian tradition regards word order in languages far from actual word order use. Such tradition is not interested in the frequency of every word order but in what the basic word order is. The basic word order is determined using parsimony criteria. In contrast, (Chapter 8) is focused on the word order minimizing a certain cost function. If real sentences minimize such a function, such word order should be the most frequent. That should be the case of SVO, which is a privileged word order in the descriptive approaches of the Chomskian tradition (Kayne, 1994; Chomsky, 1995).

- *Why human language is off the chart.*
  Chomsky's statement that *'human language is off the chart'* (Chomsky, 2002) is a natural consequence of a cascade of isolating assumptions:

  – Rejecting the meaningfulness of statistical patters (e.g. rejecting Zipf's law meaningfulness).

  – A radical division between syntax and other linguistic dimensions (e.g. Zipf's law needs semantics).

  – A radical division between performance and competence (e.g. Zipf's law needs the word frequency effect).

The basic ingredients for human language are present in non-human communication systems. Here we have provided the clues for understanding what makes language complex.

10.3 General approaches to the origins of language

There has been a long debate about whether human language can be explained from Darwinian evolution (Pinker and Bloom, 1990; Nowak, 2000b) (assuming natural selection operates at the level of syntax) or it is a side effect of another process or function (Chomsky, 1972; Chomsky, 1982b; Chomsky, 1982a; Chomsky, 1988a; Chomsky, 1988b; Gould, 1987; Chomsky, 1991). The first argument has to face a fundamental question: how can syntax be selected if there is no syntax at all? If syntax is assumed and natural selection has to make a decision about selecting or not, it is not a totally fair game, since how syntactic communication naturally appears is not explained.

We have seen that Zipf's law (with $\beta \approx 2$), a natural consequence of communication constraints, provides connectedness (a necessary condition) and linking words (Chapter 7), supporting syntax is a by-product and not the object of Darwinian selection. We are an not denying selection operates at the communication level (coding/seconding) but at the syntactic level. Syntax is not only a side-effect of referential constraints but also a consequence of cognitive pressures. If a communication system operates at the perfect communication phase (Chapter 3), increasing the number of objects forces the set of signals to grow, increasing the coding effort and other costs (Chapter 3). There is a threshold in the number of objects to describe. The effect of crossing such a threshold is what we call a referential catastrophe. If the number of objects goes to infinity, the number of signals also must go to infinity when the communication system is operating in the perfect communication phase (Chapter 3). This idea of a cognitive threshold was hypothesized by Noam Chomsky by means of philosophical enquiry (Chomsky, 1991) and recently formalized as a minimum amount of objects to describe for Darwinian selection to favour syntax (Nowak and Krakauer, 1999; Nowak, Plotkin, and Jansen, 2000).

Word binding is the by-product of complying with coder and decoder interests. It can be clearly seen what exaptations in human language are. Linking words are exaptations of referentially useless words (Chapter 7). Referentially useless words take over structural functions. This view agrees with the idea that nature is a tinkerer and not an engineer with an empty piece of paper (Jacob, 1977).

Our work also sheds light on the abruptist (or punctuated sensu (Gould and Eldredge, 1993)) versus gradualistic origins of language. Although fast changes in the transition to language in children and in the transition from pidgin to creole (the so-called creolization process (Bickerton, 1981; Bickerton, 1984; Romaine, 1992)), there is no empirical evidence that this was the case for the very origins of human language. Theoretical approaches have shown that transitions to syntactic communication should be abrupt in the context of cultural evolution (Kirby, 2000) or once certain thresholds are crossed in the context of Darwinian selection (Nowak and Krakauer, 1999; Nowak, Plotkin, and Jansen, 2000). Our approach here is the first formal approach to syntax abruptly emerging from strictly referential principles.

Punctuated approaches (e.g. Section 1.3.3 and Chapter 3) easily explain the gap between human language and (most of) non-human species (Ujhelyi, 1996). Syntax and symbolic reference are given only if connectedness is reached in the network of signal-object associations. Small changes in the average connectivity of the networks have no effect if they take place sufficiently far from the transition.

Formal approaches to the origins of language can be followed through a series of books (Hurford, Studdert-Kennedy, and Knight, 1998; Briscoe, 1999; *et al.*, 2000; Cangelosi and Parisi, 2002) and recent reviews (Kirby, 2002b; Christiansen and Kirby, 2003; Nowak, 2000b; Nowak and Komarova, 2001; Nowak, Komarova, and Niyogi, 2002). Such approaches can be classified into two major lines, i.e. cultural and biologial evolution. The former assume language is learned and the process leading to language is cultural evolution. The latter assume language is innate and the process leading leading to language is biological evolution, that is, Darwinian evolutionary framework based on iterated reproduction and selection of the best communicating individuals. The biological approaches have recently benefited from the use of rigorous mathematical tools borrowed from evolutionary dynamics, game and information theory (Nowak and Krakauer, 1999; Nowak, Krakauer, and Dress, 1999; Nowak, Plotkin, and Krakauer, 1999; Grassly, von Haeseler, and Krakauer, 2000; Nowak, Plotkin, and Jansen, 2000; Nowak, 2000a; Nowak, 2000b; Krakauer, 2001; Komarova and Nowak, 2001).

Different topics shed light on the similarities and differences of the present thesis with previous work,

- *The definition of language.* Here we assume language is a combination of both communication (successful information transfer among agents) and syntax or symbolic reference. Other approaches assume syntax and symbolic reference are dissociated, an never address the question of symbolic reference, at least least following the complex understanding of symbolic reference, and not simplistic approaches as in (Oliphant, 2002). Biological approaches and certain cultural approaches use the communication-syntax pair as the definition of language. In contrast, certain reductionist approaches neglect communication (Kirby, 2000; Kirby, 2001; Kirby and Hurford, 2001; Kirby, 2002a; Smith, Kirby, and Brighton, To appear). Their results must be cautiously interpreted. They are interesting models of language acquisition or syntax, but not of language in the strict sense. Language implies syntax, but syntax does not imply language.

- *The definition of symbolic reference.* As far as we know, no previous formal approach defines symbolic reference mathematically. A precise definition of symbolic reference is given in the present thesis for the first time (see Section 10.1).

- *The definition of syntax.* The evolutionary biology approaches defines syntax as the possibility of forming combinations of a few signals, Such simple kind of approach seem to be absent in cultural approaches, that make emphasis on various aspects of syntax such as compositionality (Batali, 1998; Kirby, 2000) and recursion (Kirby, 2002c; Kirby, 2000; Kirby, 2002a). For computational reasons, syntax is often simplified assuming a maximum sentence length (e.g. (Batali, 1998), a finite meaning space (e.g. (Batali, 1998; Kirby and Hurford, 2001)) or limited recursion or compositionality (e.g. (Batali, 1998)). Some of them make two heavy *a priori* assumptions such as assuming the existence of a device capable of parsing context-free grammars (Kirby, 2000; Kirby, 2002a). As it is said above, such models are therefore models acquisition by an agent pre-adapted to language. Defining syntax as combinations of a few signals, has nothing to say about the possibility of chaining such combinations for forming complex sentences. Here we have taken a stronger and more realistic requirement for syntax than in biological approaches, i.e. connectedness in the network of signal-object associations, but minimizing the amount of *a priori* assumptions.

- *The nature of signal object associations.* Here we assume (without specifying) the existence of a mechanism for forming signal-object associations but makes no emphasis on whether it is innate or acquired. Different cultural and biological evolution address that question (Kirby, 2002b). The results presented here are thus independent of the mechanism chosen, which provides us with a higher degree of abstraction.

- *The nature of meanings.* Here we assume a set of objects of reference, but does not deal with the nature of such objects. Different cultural and biological approaches address the question of how meaning is grounded (Kirby, 2002b).

- *How signals glue for signal.* When signals combine in syntactic communication, the evolutionary biology approaches explain why signals glue but do not explain how signals glue, e.g. the existence of links is assumed. Cultural approaches do not explain 'how' either. Using a recurrent neural network for forming strings of signals (Batali, 1998) or using a context-free grammar for modeling language (Kirby, 2000; Kirby, 2002a) links signals *a priori.* But these approaches raise methodological questions. As Bickerton points out (Bickerton, 2000), natural selection works on variation: it selectively increases the distribution of variations that are adaptative. But how could there be a variation in syntactic ability before there was syntax? If syntax is defined as 'gluing' words, how can word 'gluing' be selected before there is 'glue'? The same arguments applies to cultural evolution. How can learning constraint 'gluing' word systems before there is no a priori 'glue'? Here semantic compatibility is the natural 'glue'. Two signals are linked by a common object. The hypothesis supported here is that word 'gluing' is a side effect of Zipf's law.

- *The communication channel.* The approach here assumes errors during communication are irrelevant. While this is also the case of many culturally evolving systems (Hurdord, 2002), the evolutionary biology approaches consider errors as the essential factor for the emergence of syntactic communication.

- *The origins of consensus. Referential decoding least effort.* The evolutionary biology approaches introduce consensus in the fitness function explicitly while certain cultural approaches, such as the iterated learning model, focus on various aspects of syntax neglecting consensus (Kirby and Hurford, 2001; Smith, Kirby, and Brighton, To appear). Some approaches consider stability, that is, how similar are the language of a child agent with regard to its parent agent (Kirby and Hurford, 2001). That can be seen as an intermediate situations between requiring consensus and not requiring it. Here, consensus is assumed. The approaches here neglect the fact that communication is a social phenomenon. We could assume a population of agents with its own linguistic knowledge (e.g. signal-object associations) in order to allow for discrepancies. Our models are a sort of mean field, that is, we assume the the linguistic knowledge of every agent is (almost) the same. Luckily, different sources indicate that consensus in the signal-object associations a population is expected in innate or culturally evolved signaling systems (Kirby, 2002b). Most researchers would probably agree that evolution by natural selection can tune a simple communication system under reasonably ecological assumptions. Innate signaling systems are a mature area (Kirby, 2002b). As for culturally evolved systems, a rich array of simulation experiments reaching consensus is supported by strong results from the study of matrix of interactions between agents. Convergence to a common language is warranted for irreducible stochastic matrices of interaction (Cucker and Smale, 2002).

- *Referential coding effort.* The present thesis is the first formal study of the coding effort an its consequences.

- *The origins of syntax.* Here syntax is shown to be a side effect of communication principles in a noiseless channel (i.e. coding/decoding least effort) whereas the evolutionary biology approaches assume a Darwinian competition between syntactic an non syntactic strategies. Syntax is not a side-effect in the strict sense for the latter since how words glue is not explained.

- *Zipf's law.* The evolutionary biology approaches generally assume (but not explain) Zipf's law when signal frequencies are needed. They suggest intermittent silence models as a possible explanation. As far as we know, Zipf's law is not a studied and not even assumed in the cultural approaches to language. Here Zipf's law is the bulk of our argumentation and not assumed when studying communication systems.

- *What favours small vocabularies.* When explaining why vocabularies in non-human species are mostly small (assuming articulatory constraints are irrelevant), the evolutionary biology models hypothesize that communication with errors forces to choose a distinguishable subset of signals. Cultural approaches to the origins of language, mostly based on computer simulations, pay no attention to vocabulary size constraints, probably due to the fact that in some cases the possible vocabulary can be arbitrarily large or its limited in order to run computer simulations fast. Here, the coding effort precludes vocabulary growth.

- *The basic mechanism for language.* We assume optimization leads to syntax. For the evolutionary biology approaches Darwinian evolution is the optimization mechanism. Cultural evolution leads to syntax even when there is no population turnover (Batali, 1998; Kirby, 2002c). Different approaches where consensus is neglected lead to syntactic communication through leaning constraints (Kirby and Hurford, 2001). Here, the exact mechanism is left for future research.

- *Language is a phase or a transition between phases.* The evolutionary biology approach hypothesizes that overcoming phase transitions could have been the origin of the emergence of syntax (Nowak and Krakauer, 1999). We do not make any consideration about the consequences of increasing linking connectivity in a way that the meaning of signals degenerates. We discussed in Chapter 7 that words link through their meanings. Therefore increasing signal combinations too much can lead to referentially useless words. Therefore, a certain trade-off must be present. When different phases are identified in models of cultural evolution (Kirby, 2000), language is the final phase. Here, we show that early human communication could have benefited from remaining in a referential phase transition. We do not need an extra argument for keeping the amount of signal-object associations at the minimum value needed for connectedness Zipf's law does it by placing the system not only in the edge of a referential phase transitions but also in the edge of a connectedness phase transition.

- *Gradual versus abrupt evolution.* Here we provide support the transition to syntax must be abrupt, as well as in biological and certain cultural approaches where sharp transitions between phases are found.

- *Cognitive thresholds.* Biological approaches and the approach here show thresholds in the number of objects to describe lead to syntax but through totally different paths, i.e. a referential catastrophe here and an error threshold in the evolutionary biology approach. As far as we know, cultural approaches have not addressed the question of cognitive thresholds directly.

Chapter 11

Conclusions

11.1 A new framework for the study of human language

The evidence presented here claims for a new view of language and its origins through the eyes of physics. A phase transition to connectedness is a precise definition of the misleading formulation of a 'sudden macromutation' for the origins of syntax (Bickerton, 1990; Bates, Thal, and Marchman, 1989) (in the best case misleading because it calls for a genetic driven origins). *A brain achieving a certain level of complexity* as the cause of human language" (Chomsky, 1991) can be more accurately formalized in terms of cognitive thresholds, that is, thresholds in the number of objects to describe (Chapter 3 and Nowak, Plotkin, and Jansen (2000)). Connectedness is a necessary condition for syntax but is equivalent to a rudimentary form of syntax where recursion is not implied but allowed. Connectedness is a concept borrowed from graph theory. Connectedness assumes that the Euclidean distance between vertices does not matter. Percolation is similar to connectedness but Euclidean distance between units matters. Percolation arises from a number of problems in condensed matter physics. An example is a fluid passing through a porous media. The porous media (e.g. a rock) have many small random channels. We wonder if the fluid can flow through the material. The answer depends on the amount channels and the way they are arranged. If there are too few channels the fluid cannot pass through. It is said the system does not percolate. When there are more channels, the fluid can pass through the solid. In that case, it is said the system is percolating (Stauffer and Aharony, 1994). A rudimentary form of syntax and percolation in water passing trough a porous media are essentially the same phenomenon.

Gaps between human language and (most of) non-human species communication systems can be replaced by phase transitions. If the transition to language is a continuous phase transition (Section 1.3.3 and Chapter 3), intermediate stages should be very rare but may exist. The origins of the limited abundance of intermediate stages is due to the fact that the transition takes place in a very narrow domain of the parameter governing the transition between phase, e.g. p in Section 1.3.3. It must be understood that the situation in the dual least effort model is slightly different (Chapter 3). There, human language (Zipf's law) is not one of the phases but the edge between two phases. Zipf's law is not a phase but the transition itself. Thus, communicating (without Zipf's law) and non-communicating species are different phases.

Phase transitions can explain abrupt changes, but abrupt changes to not imply phase transitions. For instance, path dependent processes (Arthur, 1994; Jenkins, 2000) (not necessarily phase transitions) may underly the fast word order rearrangement and fixation from pidgin to creole languages (Bickerton, 1990).

The sudden emergence of language is not only a consequence of a phase transition, but also a consequence of an explosion concerning language features. We have seen that Zipf's law has non-trivial consequences, i.e. connectedness (and therefore combinatorics), linking words (e.g. prepositions) as those of human language and hierarchy (Chapter 7). The present thesis limits the scope where natural selection operates. Non-trivial features of language are given for free in agreement with S. Kauffman's understanding of evolution (Kauffman, 1993). The objects upon natural selection could operate in human language need to be reconsidered. Proponents of natural selection (Pinker and Bloom, 1990; Bates, Thal, and Marchman, 1989) in human language must first exploit deductions upon language laws such as Zipf's before wondering how a certain feature evolved from natural selection. The present work is the beginning of axiomatico-deductive approach to the origins of language. Axiomatico-deductive system statistical patterns have been previously developed in quantitative linguistics (Köhler, 1986).

Zipf's law (with $\beta \approx 2$) is called a law because of its robustness. It appears in a wide range of different linguistic contexts (Balasubrahmanyan and Naranan, 1996). But Zipf's is an empirical law and not a law in the strict sense (Li, 1998). We have shown that Zipf's law is a power function linked to a continuous phase transition. Therefore, the gate towards an analytical approximation for such a statistical regularity is open and so is the possibility of considering Zipf's law a law of reference.

There are further reasons for considering Zipf's law as a fundamental law of language. First, we have mentioned above that different syntactic features of language stem from that law. Zipf's law has a great deductive power. Second, the construction of a theory of language correlated with the evolution of language needs to establish a set of basic laws where one or more deduction steps can be mapped into evolutionary stages. For instance, Zipf's law with $\beta \approx 2$ leads to connectedness and connectedness with minimal E_D leads to Zipf's law with $\beta \approx 2$ (assuming word frequencies must be a power function). We have seen that connectedness must be regarded as a by-product and not as a cause of Zipf's law (Chapter 1). Therefore, Zipf's law should be among the set of basic laws. The present thesis puts a step forward for a theory of language.

The deep similarities among languages on Earth call for a universal grammar.

Statistical physics provides powerful techniques for quantifying such similarities. The features studied here (Chapter 6) suggest that all existent human languages belong to the same universality class but adding more languages or considering other statistical measures could lead to a deeper classification of languages than typology currently provides.

11.2 Contributions

The main theoretical contribution of the present work are:

- The existence of a core and a peripheral lexicon in a community of speakers of the same language. Zipf's law with $\beta \approx 2$ is only found in the former (Chapter 2).

- Intermittent silence models are not either models for Zipf's law or null hypothesis. Intermittent silence models neglect words are used according to their meaning (Chapter 5) and the existence of a mental lexicon.

- Word frequency distributions must be explained assuming the frequency of a word is determined by the frequency of its meanings. The simplest approach is to assume all objects have the same frequency and define signal frequency proportional to the number of objects (Chapter 4).

- Zipf's law is a characteristic power distribution of a continuous phase transition (Chapter 3).

- Zipf's law is not a mere empirical regularity but a law of reference (Section 11.1).

- Maximizing Shannon's definition of information transfer is not sufficient for modeling complex natural communication systems. Minimizing the entropy of codes (coding least effort) is a conflicting constraint needing to be considered (Chapter 3).

- The scale-free degree distributions of syntactic dependency networks are a consequence of Zipf's law for word frequencies (Chapter 7).

- The transition to syntax must be sharp and continuous at least in children (Section 1.3.3). If syntax appeared after Zipf's law in early hominids, the transition to syntax would have been as abrupt as the transition to Zipf's law.

- A formal definition of symbolic reference (Chapters 7,10).

- Connectedness is a necessary condition for syntax (Section 1.3.3 and 7) and symbolic reference (Chapter 7).

- Zipf's law with $\beta \approx 2$ is a core pattern of human language. From it follows connectedness (but not recursion), the existence of linking words and non-trivial regularities at the level of word-word syntactic interactions such as hierarchical organization and disassortative mixing (Chapter 7).
- The exponent of Zipf's law contains information about the communicative efficiency of the system (Chapter 4).
- Simple strategies for coding and decoding (Chapters 3,4) do not lead to Zipf's law, suggesting Zipf's law is the hallmark of complex communication systems.
- The role of natural selection for shaping human language need to be reconsidered because of an explosion of non-trivial language features following directly from Zipf's law.
- Human language is a by-product of referential principles.
- Syntactic dependency networks belong essentially the same universality class regardless of the language into consideration. Instead of a universal grammar hypothesis to account for world languages essential similarities, a single universality class is proposed. We have seen that disassortative mixing, hierarchical organization and a negative correlation between clustering and degree (Chapter 6) follow from Zipf's law (Chapter 7). Therefore, assuming Zipf's law is universal in world languages, there is only one possibility: the properties of syntactic dependency networks are universal. If not, there are languages where Zipf's law not obeying Zipf's law with $\beta \approx 2$ and therefore, very unlikely, there are languages where the maximum tension between hearer and speaker is avoided.
- Syntactic dependency networks and semantic networks (e.g. thesaurus networks) belong to different universality classes.
- Euclidean distance minimization explains projectivity, the distribution of the distance between syntactically linked words and SVO order privileged position in real language use.
- Projectivity is not a principle but a side effect of a Euclidean distance minimization principle (a minimum linear arrangement) (Chapter 8).
- The outcome of network distance and link density minimization is a small network morphospace.

The empirical contributions of this work are:

- Two domains in the frequency of words in multi-author collections of texts (Chapter 2).
- Words of the same length also obey Zipf's law with $\beta = 2$ (Chapter 5 and Appendix C). In contrast, intermittent silence fails, even using biased real letter probabilities (Appendix C).

- Nouns in multiauthor corpora follow Zipf's law with $\beta = 3.35$ (Chapter 4).

- Maximum entropy in p_k (the proportion of signals with k objects) is found in the vicinities of a referential phase transition (Chapter 4).

- Most of syntactically related words in sentences are at distance less or equal than two (more precisely, 87%; Chapter 8).

- The distance between syntactically related words takes an exponential form (Chapter 8).

- Global syntactic dependency networks exhibit power functions in their degree distributions and betweenness centrality distributions (Chapter 6).

- Sentence syntactic dependency networks exhibit a rather exponential like degree distribution (Chapter 6).

- Global syntactic dependency networks are *small-worlds* and show disassortative mixing and hierarchical organization (Chapter 6).

- Syntactic dependency degree is a linear function of word frequency (Chapter 6).

Appendix A

Data sources

The British National Corpus

An English corpus formed by a collection of text samples (generally not longer than $45,000$ words). It is synchronic (it includes imaginative texts from 1960, informative texts from 1975), general (not specifically restricted to any particular subject field, register or genre), monolingual (it comprises text samples which are substantially the product of speakers of British English) and mixed (it contains both examples of both spoken and written English). Additional information is available at http://info.ox.ac.uk/bnc.

The Czech Academy Corpus

A Czech dependency corpus annotated by Ludmila Uhlířová, Jan Králík among others (Uhlírova, Nebeská, and Králík, 1982; Těšitelová, 1985). The corpus was compiled at the Czech Language Institute, Prague, within 1970-1985.

The corpus contains 562820 words and 31701 sentences. Many sentence structures are incomplete in this (i.e. they have less than $n-1$ links, where n is the length of the sentence). The proportion of links provided with regard to the theoretical maximum is about 0.65.

The Dependency Grammar Annotator samples

The Romanian corpus formed by all sample sentences in the Dependency Grammar Annotator website [1]. It contains 21275 words and 2340 sentences. The syntactic annotation was performed by hand.

The Negra Corpus

A German corpus (The Negra Corpus 1.0) containing 153007 words and 10027 sentences. The formalism used is based on the phrase structure grammar.

[1] http://phobos.cs.unibuc.ro/roric/DGA/dga.html

Nonetheless, for certain constituents, the head word is indicated. Only the head modifier links between words at the same level of the derivation tree were collected. The syntactic annotation was performed automatically. The proportion of links provided with regard to the theoretical maximum is about 0.16.

Appendix B

Conversion between power distributions

Frequency distributions can be presented in terms of frequency versus rank of as a probability (density) function. If the distribution is follows a plain power function, as in Zipf's law, the frequency versus rank representation is

$$P(i) = p_1 i^{-\alpha} \tag{B.1}$$

where $P(i)$ is the normalized frequency of the i-th most frequent word in the sample, $i > 0$, $\alpha \approx 1$ (Zipf, 1972a; Casti, 1995; Tsonis, Schultz, and Tsonis, 1997) and p_1 is the probability of the most frequent word (Tuldava, 1996).

The same power distribution can also be presented as probability (density) function:

$$P(f) \propto f^{-\beta} \tag{B.2}$$

where $P(f)$ is the probability that a word is present f times in a text.

We can relate the rank with the probability density function. Let us denote by $m_n = TQ(n)$ the number of words having population n, where T is the total number of word in the sample. Then, the rank is given by

$$R(n) = \int_n^\infty m_{n'} dn' \tag{B.3}$$

and the most frequent word has $R = 1$, the second most frequent word has $R = 2$, and so on, for decreasing values of n in the integral. Eq. B.3 establishes a general relation between the rank of an event in the sample and the probability distribution according to the event frequency. Substituting $R \propto n^{-1/\alpha}$ (obtained from Eq. B.1) and Eq. B.2 in Eq. B.3 we immediately get $n^{1-\beta} \simeq n^{-1/\alpha}$, from where

$$\alpha = \frac{1}{\beta - 1} \tag{B.4}$$

$$\beta = \frac{1}{\alpha} + 1 \tag{B.5}$$

If $\alpha = 1$ then β should be 2.

Appendix C

Intermittent silence in depth

One obvious question raised by the ubiquity of Zipf's law with $\beta \approx 2$ is: is it the result of some non trivial causal process? Any observed regularity in nature needs first to be studied by means of null models. One possible explanation of the Zipf's law comes from a purely random process. An early argument against any special causal explanation beyond randomness was the discovery that random sequences of letters (in which the blank space was among them) reproduced the $\alpha = 1$ exponent of words (Miller, 1957). Assume that the keys of a typewriter are typed at random. If the blank space is hit with probability q and one of the N possible letters are hit with probability $(1-q)/N$, having all letters the same probability, the distribution of words limited by blank spaces can be shown to obey Eq. 1.2 (Miller, 1957; Mandelbrot, 1966; Li, 1992).

To some extent, it has been concluded that Zipf's law does not tell anything (relevant) about language (Wolfram, 2002; Nowak, 2000a; Li, 1992; Miller and Chomsky, 1963).

Such a conclusion comes, in our view, from the misleading comparison between rank distributions. When the lexical spectrum is plotted for the monkey language, the differences between random and non random sequences become dramatic. Fig. C.1 shows the normalized frequency versus rank and the lexical spectrum for a monkey language with $q = 0.18$ and $N = 26$. The former shows $\alpha \approx 1$. The later should show an exponent $\beta \approx 2$ as predicted by Eq. 1.3 but no power domain consistent with Zipf's law (with $\beta \approx 2$) can be identified and it differs greatly from its counterpart in Fig. 1.1. It is tempting to think the statistical structure of both distributions is completely different.

Vocabulary growth in intermittent silence models is faster than in real texts (Cohen, Mantegna, and Havlin, 1997). $N = 26$ leads to $11,881,376$ different 5-letter words, far from the about $1,7$ million words of intermittent silence text in Fig C.1. If sampling effects are responsible for the surprising plot in Fig. C.1 B, the lexical spectrum with $N = 2$ should improve (there are only 32 different

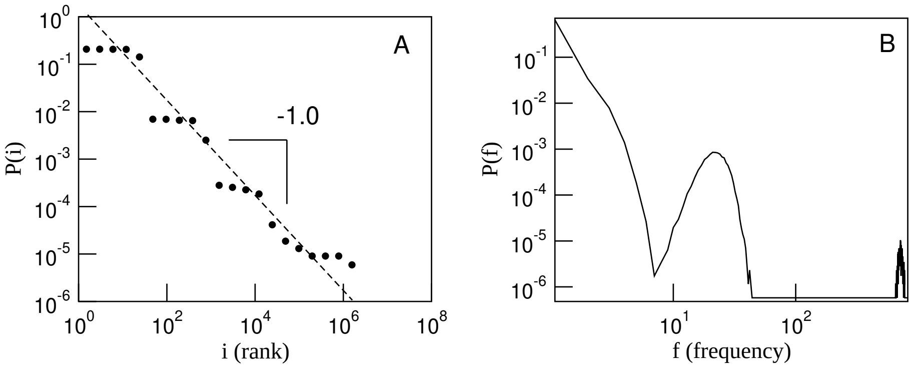

Figure C.1: Frequency versus rank (A) and lexical spectrum (B) of intermittent silence text formed by $1,731,411$ different words ($4 \cdot 10^6$ total words). The alphabet has $N = 26$ letters (all having the same probability) and the probability of blank space is $q = 0.18$. The exponent in (A) is $\alpha = 1$ while no power law with $\alpha \approx 1$ seems to fit in (B).

5-letter words). Fig. C.2 shows that not only the frequency versus rank plot improves but also the lexical spectrum. Nonetheless, the quality of the latter is still clearly lower than that of a real text. The analytically predicted exponents are obviously valid in Fig. C.1-C.2 B but intermittent silence models like these reveal high sampling sensitiveness when compared to real texts.

It might be thought the monkey language we have employed is simplistic. All letters have the same probability, which is not realistic. If an intermittent silence text is generated with letter probabilities obtained from Moby Dick, the frequency versus rank plot loses the step-like appearance (solid line in Fig. C.3 A) while the lexical spectrum improves (Fig. C.3 B). Notice that the improvement can not be attributed to a smaller vocabulary (about 1.7 million word in the unbiased case) but a less restrictive way of filling the spectrum.

An additional source of disagreement comes from the analysis of word distributions of a certain length. Monkey languages imply word length follows an exponential distribution given by

$$P(L) \propto (1-q)^L \tag{C.1}$$

where $P(L)$ is the probability of words formed by L letters. In contrast, word length is modeled with log-normal (Balasubrahmanyan and Naranan, 1996; Naranan and Balasubrahmanyan, 1992b) or Poissonian distributions (Wimmer et al., 1994). Empirical studies show that there is a typical length $L > 1$ and long tails may appear. If all letters have the same probability, monkey languages predict that words having the same length have the same frequency. The dashed line in Fig. 1.1 A shows the distribution of words in Melville's Moby Dick having the same length, which is clearly Zipf's with $\alpha \approx 1$. In contrast, the equivalent in a monkey language in which all letters have the same probability is a uniform

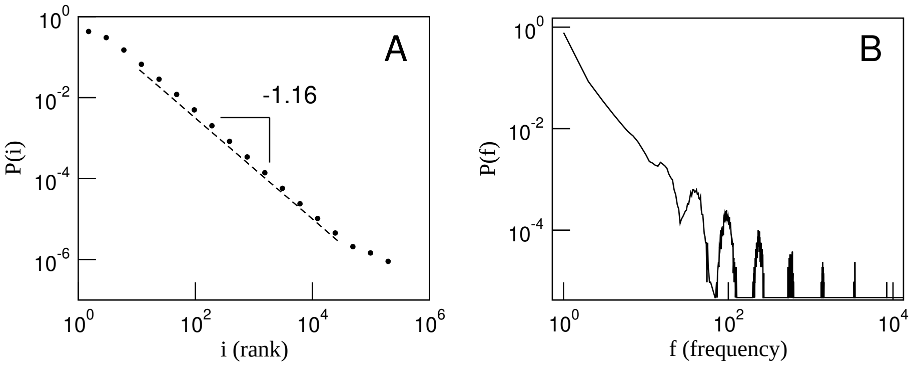

Figure C.2: Frequency versus rank (A) and lexical spectrum (B) of intermittent silence text formed by 212, 197 different words ($4{\cdot}10^6$ total words). The alphabet has $N = 2$ letters (all having the same probability) and the probability of blank space is $q = 0.18$. The exponent in (A) is $\alpha = 1$ and the quality of the lexical spectrum it higher than with $N = 26$.

distribution and the distribution of letters of a monkey language with realistic letter frequencies is the dashed line in Fig. C.3 A. Both are clearly far from a power function with $\alpha \approx 1$.

In a previous study (Lepold, 1998), it has been shown that English words of length 5 follow Zipf's law with $0.8 < \beta < 0.9$, apparently contradicting the exponent close to $\beta = 2$ found for the British National Corpus in Fig. 1.1 ($\alpha = \frac{1}{\beta-1} = 1.03 \pm 0.005$). Furthermore, such a previous study shows that β takes values in $[0.8, 1.4]$ for length 5 in English, Dutch, Finnish, German and Polish. Such deviations may be explained by the more reduced corpora from which word frequencies were extracted with regard to the British National Corpus and the possibility that words outside the core lexicon with $\beta \approx 3/2$ (Chapter 2) are lowering the exponent. The calculations in Fig. 1.1 come from a single author text, whereas the corpora in (Lepold, 1998) are multiauthor. To sum up, Zipf's law with $\beta \approx 2$ is recovered for the typical length in single author corpora. The possibility that Zipf's law with $\beta \approx 2$ appears independently of the length class under consideration can not be denied. Since word length is correlated with word frequency (Best and Zhu, 1994; Wimmer et al., 1994), the longer the word, the higher the probability it is not a core word. Therefore, such independence should be more clearly observed for the shortest words.

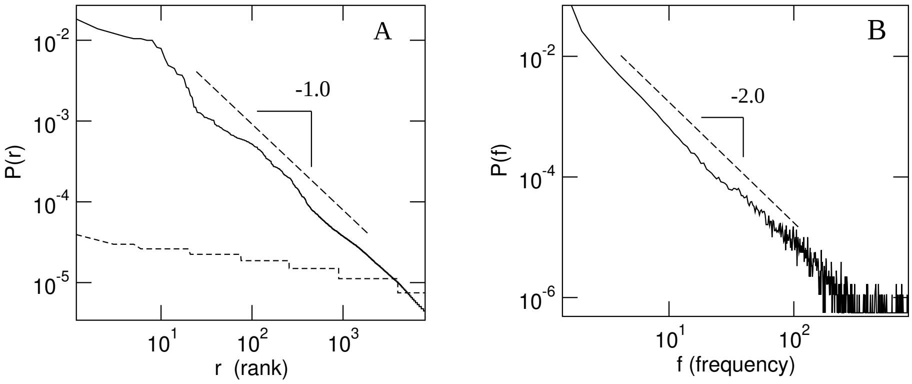

Figure C.3: Frequency versus rank (A) and lexical spectrum (B) of a monkey language formed by $1,795,617$ different words ($4 \cdot 10^6$ total words). Character probabilities were obtained from Melville's Moby Dick. The dashed line in A shows the frequency versus rank for words having length 5, which is the average length of words in Melville's book. The intermittent silence text has $238,891$ different 5-letter words. The exponent in (A) is $\alpha = 1$ while the slope in (B) is $\alpha = 2.0$.

Appendix D

Simon's model

Appendix

Here we consider we calculate the exponent of the Simon's model that is simpler than the provided by Simon in (Simon, 1955).

Lets consider a generalized Simon process in which the urn contains m_0 words at the 0-th step and m identical words are added to the urn at every time step t ($t > 0$). The words added are new with probability ψ or they are a copy of an existing word (which is chosen at random from the urn) with probability $1-\psi$. The exponent of the frequency distribution can be easily derived with the mean field schema used for Barabási-Albert scale-free network model (Barabási, Albert, and Jeong, 1999). If k_i, the number of occurrences of the i-th word, and also t are considered to be continuous, then the expected variation of k_i is

$$\frac{dk_i}{dt} = (1-\psi)m\pi_i \tag{D.1}$$

Substituting $\pi_i = \frac{k_i}{\sum k_i}$ and $\sum k_i = m_0 + mt$ into Eq. D.1 we obtain

$$\frac{dk_i}{dt} = \frac{(1-\psi)mk_i}{m_0 + mt} \tag{D.2}$$

Integrating Eq. D.2 for a word that appeared for the first time at $t = t_i$ with m copies leads to

$$k_i = m\left(\frac{m_0 + mt}{m_0 + mt_i}\right)^{(1-\psi)m} \tag{D.3}$$

From Eq. D.3 it follows that

$$P(k_i < k) = P(t_i > A)$$

where

$$A = \frac{1}{m}\left(\frac{m^{\frac{1}{(1-\psi)m}}}{k^{\frac{1}{(1-\psi)m}}} - m_0\right) \tag{D.4}$$

Thereafter

$$P(k_i < k) = 1 - P(t_i \leq A) = 1 - \int_0^A P(t_i) dt_i \qquad \text{(D.5)}$$

Substituting $P(t_i) = \frac{1}{m_0+mt}$ into Eq. D.5 and being $P(t_i)$ independent of t_i if follows that

$$P(k_i < k) = 1 - \frac{1}{m_0 + mt} A \qquad \text{(D.6)}$$

Substituting D.6 into $P(k) = \frac{dP(k_i<k)}{dk}$ we obtain

$$P(k) = \frac{m^{\frac{1}{(1-\psi)m}-2} m_0 + mt}{1-\psi} \frac{1}{k^{\frac{1}{(1-\psi)m}+1}} \qquad \text{(D.7)}$$

and hence

$$P(k) \propto k^{-1-1/(1-\psi)m}$$

For $m = 1$

$$P(k) \propto k^{-1-1/(1-\psi)}$$

as it is expected for the basic Simon process.

Appendix E

Basic network topologies

Throughout Chapter 9, different trivial topologies appear. Table E.1 summarizes their features indicating the value of λ at which they appear. Although Chapter 9 is concerned with what happens for $\lambda \geq 0$, notice that the linear graph is the expected outcome for $\lambda < 0$, since it implies distance maximization and density minimization. The remaining of this section is devoted to proof that a linear graph and a star graph have the maximum finite distance and the minimum distance (with the constraints of connectedness and having the smallest amount of edges).

A linear graph is a graph having the maximum finite distance or in other words, it is the connected graph having the maximum distance. We will proof it through induction on n. For $n = 2$, there is only one possible connected graph, which trivially has the maximum distance. All linear graphs having the same amount of vertices have the same average vertex-vertex distance. If the graph in Eq. 9.2 has the maximum distance for n vertices, will it still be the longest for $n+1$ vertices? Assuming that the graph in Eq. 9.2 is the longest for n vertices, the longest graph of $n + 1$ vertices has to be formed by the longest graph of order n and a new a vertex linked to one of the n existing vertices. Here we

Topology	ρ	D	C	H	λ
Poisson	ρ	$\approx \frac{logn}{log(\rho(n-1))}$	ρ	$-$	0
Star	$2/n$	$\frac{2(n-1)}{n}$	0	$logn - \frac{(n-1)}{n}log(n-1)$	$-$
Complete	1	1	1	0	1
Linear	$2/n$	$\frac{n+1}{3}$	0	$\frac{1}{n}((n-2)log(n-2) + 2log2)$ $-logn$	$\lambda < 0$

Table E.1: Different trivial topologies with density (i.e. normalized amount of links) ρ, average vertex-vertex distance D, clustering coefficient C, degree distribution entropy H and the values of λ where they are optimal. $-$ indicates absence of known analytical result.

define the total vertex-vertex distance as

$$D_n = \sum_{i<j} D_n(i,j) \tag{E.1}$$

where $D_n(i,j)$ is the minimum distance from the i-th vertex to the j-th vertex. We define the average vertex vertex distance as

$$< D_n >= D_n / \binom{n}{2}$$

If D^k_{n+1} is the contribution to D_{n+1} when the new vertex is linked to the k-th existing vertex, $1 \leq k \leq n$, such an $n+1$-vertex graph obeys

$$D_{n+1} = D_n + D^k_{n+1} \tag{E.2}$$

where

$$D^k_{n+1} = \sum_{i=1}^{k} i + \sum_{i=2}^{n-k+1} i$$

Previous equation leads to

$$D^k_{n+1} = \binom{n+1}{2} \tag{E.3}$$

for $k = 1$ and $k = n$. In general,

$$D^k_{n+1} = k^2 - (n+1)k + \frac{n^2+3n}{2}$$

D^k_{n+1} has one single non-assymptotical minimum (at $k^* = (n-1)/2$) and no non-assymptotical maximum so D^k_n is maximal for $k = 1$ and $k = n$ and $1 \leq k \leq n$. $k = 1$ or $k = n$ correspond to a graph order $n+1$ satisfying Eq. 9.2, as we wanted to proof.

Substituting Eq. E.3 into E.2, we get the longest graph of order n satisfies

$$D_n = D_{n-1} + \binom{n}{2}$$

Expanding the previous recursion we get

$$D_n = \sum_{i=2}^{n} \binom{n}{2} = \frac{1}{2}\left(\sum_{i=1}^{n} i^2 - \sum_{i=1}^{n} i\right)$$

After some algebra we have $D_n = n(n^2-1)/6$ and thus $< D_n >= (n+1)/3$

It can also be shown through induction on n that a star graph with a degree distribution

$$p_k = \frac{n-1}{n}\delta_{k,1} + \frac{1}{n}\delta_{k,n-1} \tag{E.4}$$

has the minimum distance possible among all possible graphs having $n-1$ links. For $n = 2$, the only connected graph (and thus the only with finite distance) trivially is the best one having $n-1$ links. If we assume that the graph described in Eq. E.4 is the optimal for n vertices, the optimal graph of $n+1$ vertices has $d_{n+1} = d_n + \Delta^k_{n+1}$ where Δ^k_{n+1} is the contribution to D_{n+1} of the new vertex when linked to the k-th existing vertex. Thereafter, $\Delta^1_{n+1} = 2n - 1$ and $\Delta^k_{n+1} = 3(n-1)$ for $1 < k \leq n$. $\Delta^1_{n+1} < \Delta^{k>1}_{n+1}$ holds for $n > 2$, so the graph of order $n+1$ obeying Eq. E.4 is also the best one with $n-1$ links.

Appendix F

The distance between linked words

p_d, the probability that two linked words are at distance d, can be derived using the minimum entropy principle (Kapur, 1989a). Knowing that the prior distribution is $\mathcal{P}(d) = \frac{2n-d}{n(n-1)}$ and assuming there is no distance minimization, we may define the following functional

$$E = H_B - \alpha \sum_{d=1}^{n-1} p_d$$

where H_B is the Bayesian entropy defined as

$$H_B = -\sum_{d=1}^{n-1} p_d log \frac{p_d}{\mathcal{P}_d}.$$

$\frac{\partial E}{\partial p_d} = 0$ leads to

$$p_d = \mathcal{P}_d e^{-1-\alpha}.$$

The constraint $\sum_{d=1}^{n-1} p_d = 1$ gives $p_d = \mathcal{P}_d$ as expected.

Assuming $< d >= \sum_{d=1}^{n-1} dp_d$, the average distance between linked words is minimized, we may define the functional

$$E = H_B - \alpha \sum_{d=1}^{n-1} p_d - \beta \sum_{d=1}^{n-1} dp_d.$$

Thus, $\frac{\partial E}{\partial p_d} = 0$ leads to

$$p_d = \mathcal{P}_d e^{-1-\alpha-\beta d}$$

which we may write as

$$p_d = a(n-d)e^{-\beta d}$$

with

$$a = \frac{2e^{-1-\alpha}}{n(n-1)}.$$

The constraint

$$\sum_{d=1}^{n-1} p_d = 1$$

leads to

$$a = \left(\sum_{d=1}^{n-1} (n-d)e^{-\beta d} \right)^{-1}. \tag{F.1}$$

The constraint

$$\sum_{d=1}^{n-1} dp_d =< d >$$

leads to

$$a = \frac{< d >}{\sum_{d=1}^{n-1} d(n-d)e^{-\beta d}} \tag{F.2}$$

Minimizing the function

$$F = (ab- < d >)^2$$

with

$$b = \sum_{d=1}^{n-1} d(n-d)e^{-\beta d}$$

we may obtain the value(s) of β. Knowing

$$\int_0^\infty x^n e^{-ax} dx = \frac{\Gamma(n+1)}{a^{n+1}}$$

we may write Eq. F.1 as

$$a \approx \left(\frac{n\Gamma(1)}{\beta^2} - \frac{\Gamma(2)}{\beta^3} \right)^{-1} \tag{F.3}$$

and Eq. F.2 as

$$a \approx \frac{< d >}{\frac{n\Gamma(2)}{\beta^2} - \frac{\Gamma(3)}{\beta^3}} \tag{F.4}$$

for large n. Right sides of Eq. F.3 and F.4 together give

$$< d > n\beta^2 - (< d > +n)\beta + 2 \approx 0. \tag{F.5}$$

Thereafter, we have

$$\beta \approx \frac{n- < d > \pm(< d >^2 -6n < d > +n^2)^{1/2}}{2 < d > n}.$$

Appendix G

Related publications

Chapter 2 is based on (Ferrer i Cancho and Solé, 2001). Chapter 3 is based on (Ferrer i Cancho and V. Solé, 2003). Chapter 4 is based on (Ferrer i Cancho, 2003). Chapter 6 is based on (Ferrer i Cancho, Solé, and Köhler, 2003). Chapter 7 is based on (Ferrer i Cancho, Bollobás, and Riordan, 2003). Chapter 9 is base on (Ferrer i Cancho and Solé, 2003). An updated list of the publications in the present thesis can be found in

http://complex.upf.es/index.php?page=4&subpage=2&author=Ramon+Ferrer.

List of Figures

List of Tables

References

Adamic, L. A. 1999. The small world web. *Procedings of the ECDL'99 Conference, LNCS 1696, Springer,*, pages 443–452.

Akmajian, Adrian. 1995. *Linguistics. An Introduction to Language and Communication.* MIT Press. Chapter 2.

Albert, Réka, Hawoong Jeong, and Albert-László Barabási. 2000. Error and attack tolerance of complex networks. *Nature*, 406:378–381, July.

Altmann, Gabriel. 1978. Towards a theory of language. *Glottometrika*, 1:1–25.

Altmann, Gabriel. 1993. Science and linguistics. In Reinhard Köhler and Burghard Rieger, editors, *Contributions to Quantitative Linguistics.* Kluwer, Dordrecht, Boston, London, pages 3–10.

Amaral, Luis A. Nunes, Antonio Scala, Marc Barthélémy, and H. Eugene Stanley. 2000. Classes of behaviour of small-world networks. *Proc. Natl. Acad. Sci.*, 97(21):11149–11152, October.

Antos, A. and I. Kontoyiannis. 2001. Convergence properties of functional estimates for discrete distributions. *Random Structures & Algorithms*, 19:163–193.

Arends, Muysken, and Smith, editors, 1995. *Pidgins and creoles: an introduction*, pages 33–35. Benjamins, Amsterdam.

Arthur, W. Brian. 1994. *Increasing returns and path dependence in the economy.* The University of Michigan Press, Michigan.

Ash, Robert B. 1965. *Information Theory.* John Wiley & Sons, New York.

Baayen, R. Harald. 2001. *Word frequency distributions.* Kluwer Academic Publishers, Dordrecht.

Babyonyshev, M. and E. Gibson. 1999. The complexity of nested structures in Japanese. *Language*, 75:423–450.

Balasubrahmanyan, V. K. and S. Naranan. 1996. Quantitative linguistics and complex system studies. *J. Quantitative Linguistics*, 3(3):177–228.

Balasubrahmanyan, V. K. and S. Naranan. 2000. Information theory and algorithmic complexity: applications to linguistic discourses and DNA sequences as complex systems. part ii: Complexity of DNA sequences, analogy with linguistic discourses. *J. Quantitative Linguistics*, 7(2):153–183.

Ball, Phillip. 2003. Language evolved in a leap. *Nature*, January. Science update.

Banavar, Jayanth R., Amos Maritan, and Andrea Rinaldo. 1999. Size and form in efficient transportation networks. *Nature*, 399:130–132.

Barabási, Albert-László and Réka Albert. 1999. Emergence of scaling in random networks. *Science*, 286:509–511, October.

Barabási, Albert-László and Réka Albert. 2002. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97.

Barabási, Albert-László, Réka Albert, and Hawoong Jeong. 1999. Mean-field theory for scale-free random networks. *Physica A*, 272:173–187.

Barthélemy, Marc. 2003. Comment on: 'universal behavior of load distribution in scale-free networks'. *cond-mat/0304314*.

Bashkirov, A. G. 2003. On the Renyi entropy, Boltzmann principle, Levy and power-law distributions and Renyi parameter. cond-mat/0211685.

Bashkirov, A. G. and A. V. Vityazev. 2000. Information entropy and power-law distribution for chaotic systems. *Physica A*, 277:136–145.

Batali, John. 1998. Computational simulations of the origins of grammar. In J. R. Hurford, M. Studdert-Kennedy, and C. Knight, editors, *Approaches to the evolution of language: social and cognitive bases*. Cambridge University Press, Cambridge, pages 405–426.

Bates, E., D. Thal, and V. Marchman. 1989. Symbols and syntax: a darwinian approach to language development. In N. Krasnegor, D. Rumbaugh, M. Studdert-Kennedy, and R. Schiefelbusch, editors, *The biological foundations of language development*. Oxford University Press, Oxford, pages 29–65.

Best, K.-H. and J. Zhu. 1994. Zur häufigkeit von wortlänguen in texten deutscher kurzprosa (mit einem ausblick auf das chinseische). In U. Klenk, editor, *Computatio Linguae II (ZDL Beiheft 83)*. Franz Steiner Verlag, Stuttgart, pages 19–30.

Bickerton, D. 1981. *Roots of language*. Karoma Press, Ann Arbor.

Bickerton, D. 1984. The language bioprogram hypothesis. *Brain Sciences*, 7:173–222.

Bickerton, D. 1990. *Language and species*. Chicago University Press.

Bickerton, D. 1996. Catastrophic evolution: the case for a single step from a protolanguage to full human language. In M. Hurdford, J. Studdert-Kennedy and C. Knight, editors, *Approaches to the evolution of language*. Cambridge University Press, Cambridge, pages 341–358.

Bickerton, D. 2000. How protolanguage became language. In C.Knight *et al.*, editor, *The evolutionary emergence of language*. Cambridge University Press, Cambridge, pages 264–284.

Binney, J.J., N.J. Dowrick, A.J. Fisher, and M.E.J. Newman. 1992. *The theory of critical phenomena. An introduction to the renormalization group*. Oxford University Press, New York.

Bollobás, Béla. 1998. *Modern graph theory*. Graduate Texts in Mathematics. Springer, New York.

Bollobás, Béla. 2001. *Random Graphs.* Cambridge Studies in Advanced Mathematics, vol. 73. Cambridge University Press. 2nd Edition.

Bornholdt, Stefan and Kim Sneppen. 2000. Robustness as an evolutionary principle. *Proc. R. Soc. Lond. B*, 267:2281–2286.

Braitenberg, V. and A. Schuz. 1992. Basic features of cortical connectivity and some considerations on language. In J. Wind, B. Chiarelli, B. H. Bichakjian, A. Nocentini, and A. Jonker, editors, *Language origin: a multidisciplinary approach.* Kluwer, Amsterdam.

Brandes, Ulrik. 2001. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25:163–177.

Brandon, R. N. and N. Hornstein. 1986. From icons to symbols: some speculations on the origin of language. *Biol. Phil.*, 1:169–189.

Briscoe, Ted, editor. 1999. *Linguistic evolution through language acquisition.* Cambridge University Press, Cambridge.

Brown, J. H. and G. B. West, editors. 2000. *Scaling in Biology.* Oxford U. Press, New York.

Burgos, Javier D. 1996. Fractal representation of the immune b cell repertoire. *BioSystems*, 39:19–24.

Burgos, Javier D. and Pedro Moreno-Tovar. 1996. Zipf-scaling behavior in the immune system. *BioSystems*, 39:227–232.

Caldarelli, G., R. Marchetti, and L. Pietronero. 2000. The fractal properties of internet. *Europhys. Lett.*, 52:386–391.

Cangelosi, Angelo, Alberto Greco, and Stevan Harnad. 2002. Symbol grounding and the symbolic theft hypothesis. In Angelo Cangelosi and Domenico Parisi, editors, *Simulating the Evolution of Language.* Springer Verlag, London.

Cangelosi, Angelo and Domenico Parisi, editors. 2002. *Simulating the Evolution of Language.* Springer Verlag, London.

Carroll, David W. 1994. *Psychology of language.* Brooks/Cole Publishing Company, Pacific Grove, California.

Casti, John L. 1995. Bell curves and monkey languages. *Complexity*, 1(1).

Chaikin, P. M. and T. C. Lubensky. 1995. *Principles of condensed matter physics.* Cambridege University Press, Cambridge.

Cherniak, C. 1995. Neural component placement. *Trends Neurosci.*, 18:522–527.

Chitashvili, R. J. and R. H. Baayen. 1993. Word frequency distributions. In G. Altmann and L. Hřebíček, editors, *Quantitative Text Analysis*. Wissenschaftlicher Verlag Trier, Trier, pages 54–135.

Chomsky, N. 1991. Linguistics and cognitive science: problems and mysteries. In Asa Kasher, editor, *The Chomskyan turn: generative linguistics, phylosophy, mathematics and psychology*. Blackwell, Oxford, pages 26–55.

Chomsky, Noam. 1957. *Syntactic Structures*. Mouton.

Chomsky, Noam. 1965a. *Aspects of the theory of syntax*. MIT Press, Cambridge, MA.

Chomsky, Noam. 1965b. *Aspects of the theory of syntax*. MIT Press, Cambridge, MA.

Chomsky, Noam. 1972. *Language and mind*. Harcourt.

Chomsky, Noam. 1982a. Discussion of putnam's comments. In M. Piattelli-Palmarini, editor, *Language and learning: the debate between Jean Piaget and Noam Chomsky*. Harvard University Press.

Chomsky, Noam. 1982b. *Noam Chomsky and the generative enterprise: a discussion with Riny Junybregts and Henk van Riemsdijk*. Foris.

Chomsky, Noam. 1988a. *Language and problems of knowledge: the Managua lectures*. MIT Press, Cambridge MA.

Chomsky, Noam. 1988b. Prospects for the study of language and mind. Presented at the conference *Linguistics and cognitive science: Problems and mysteries", University of Tel Aviv*.

Chomsky, Noam. 1995. *The Minimalist Program*. MIT Press.

Chomsky, Noam, 2002. *On nature and language*, chapter An interview on minimalism, pages 144–146. Cambridge University Press, Cambridge, UK.

Christiansen, Morten H. and Simon Kirby. 2003. Language evolution: consensus and controversies. *TRENDS in Cognitive Sciences*, 7(7):300–307.

Chung, F. R. K. 1984. On optimal linear arrangements of trees. *Comp. & Mahts. with Appls.*, 10(1):43–60.

Cohen, A., R. N. Mantegna, and S. Havlin. 1997. Numerical analysis of word frequencies in artificial and natural language texts. *Fractals*, 5(1):95–104.

Condon, E. V. 1928. Statistics of vocabulary. *Science*, 67(1733):300.

Conner, Richard N. 1985. Vocalizations of common ravens in Virginia. *The Condor*, 87:379–388.

Corballis, M. 1991. *The lopsided ape*. Oxford University Press, New York.

Cover, T. M. and J. A. Thomas. 1991. *Elements of information theory.* Wiley, New York.

Croft, William. 1990. *Typology and Universals.* Cambridge University Press, Cambridge.

Crystal, David. 1997. *The Cambridge Encyclopedia of language.* Cambridge University Press, Cambridge, UK.

Cucker, Felipe and Steve Smale. 2002. On the mathematical foundations of learning. *Bull. Amer. Math. Soc.*, 39:1–49.

Davison, Iain. 1997. The evolution of language: assessing the evidence from nonhuman primates. *Evolution of communication*, 1(1):157–179.

de Waal, F. B. M. 1988. The communicative repertoire of captive bonobos, *pan paniscus*, compared to that of chimpanzees. *Behaviour*, 106:183–251.

Deacon, Terrence W. 1997. *The Symbolic Species: the Co-evolution of Language and the Brain.* W. W. Norton & Company, New York.

Denisov, S. 1997. Fractal binary sequences: Tsallis thermodynamics and the Zipf's law. *Phys. Lett. A*, 235:447–451.

Derbyshire, Desmond C. 1977. Word order universals and the existence of ovs languages. *Linguistic Inquiry*, 8:590–598.

de Saussure, F. 1916. *Cours de linguistic générale.* Payot, Paris.

Dewey, G. 1923. *Relative frequencies of English speech sounds.* Cambridge University Press, Cambridge, MA.

Díaz, Josep, Jordi Petit, and Maria Serna. 2002. A survey of graph layout problems. *ACM Computing surveys*, 34:313–356.

Donald, M. 1991. *Origins of the modern mind.* Cambridge University Press, Cambridge, MA.

Donald, M. 1998. Mimesis and the executive suit: missing links in language evolution. In J. R. Hurford, M. Studdert-Kennedy, and C. Knight, editors, *Approaches to the evolution of language: social and cognitive bases.* Cambridge University Press, Cambridge, pages 44–67.

Dorogovtsev, S. N. and J. F. F. Mendes. 2001. Language as an evolving word web. *Proc. R. Soc. Lond.*, 268:2595–2602.

Dorogovtsev, S. N. and J. F. F. Mendes. 2002. Evolution of random networks. *Adv. Phys.*, 51:1079–1187.

Dorogovtsev, S. N. and J. F. F. Mendes. 2003. *Evolution of networks. From biological nets to Internet and WWW.* Oxford University Press, Oxford.

Dryer, Matthew S. 1989. Large linguistic areas and language sampling. *Studies in Language*, 13:257–292.

Easton, Valerie J. and John H. McColl. web page. Statistics glossary. *http://www.cas.lancs.ac.uk/glossary_v1.1/hyptest.html*.

Ebeling, W. and T. Pöschel. 1994. Entropy and long-range correlations in literary english. *Europhysics Letters*, 26(4):241–246, May.

Ellis, Stephen R. and Robert J. Hitchcock. 1986. The emergence of Zipf's law: spontaneous encoding by users of a command language. *IEEE Trans. Syst. Man Cyber.*, 16(3):423–427.

Estoup, J. B. 1916. *Gammes stenographique*. Gauthier-Villars, Paris.

et al., C.Knight, editor. 2000. *The evolutionary emergence of language*. Cambridge University Press, Cambridge.

Fell, D. and A. Wagner. 2000. The small-world of metabolism. *Nature Biotech.*, 18:1121–1122.

Ficken, Jack P. Hailman. Millicent S. and Robert W. Ficken. 185. The 'chick-a-dee' calls of *parus atricapillus*: a recombinant system of animal communication compared with written english. *Semiotica*, 56:121–224.

Ficken, Millicent Sigler, Elizabeth D. Hailman, and Jack P. Hailman. 1994. The chick-a-dee call system of the Mexican chickadee. *Condor*, 96:70–82.

Frisch, K. Von. 1967. *The dance language and orientation of bees*. Harvard University Press, Cambridge.

Furusawa, Chikara and Kunihiko Kaneko. 2003. Zipf's law in gene expression. *Physical Review Letters*, 90:088102.

Gardner, Beatrix T. and R. Allen Gardner. 1994. Development of phrases in the utterances of children and cross-fostered chimpanzees. In *The ethological roots of culture*. Kluwer Academic Press, Netherlands, pages 223–255.

Gardner, Tim, G. Cecchi, and M. Magnasco. 2001. Simple motor gestures for birdsons. *Physical Review Letters*, 87:208101.

Garey, M. R. and D. S. Johnson. 1979. *Computers and intractability: a guide to the theory of NP-completeness*. W. M. Freeman, San Francisco.

Geissmann, T. 2000. Gibbon songs and human music from an evolutionary perspective. In N. L. Wallin, B. Merker, and S. Brown, editors, *The origins of music*. MIT Press, Cambridge MA, pages 103–123.

Gernsbacher, Morton Ann, editor. 1994. *Handbook of Psycholinguistics*. Academic Press, San Diego.

Gibson, E. 2000. The dependency locality theory: a distance-based theory of linguistic complexity. In *Image, language, brain.* The MIT Press, Cambridge, MA, pages 95–126.

Givón, T. 1979. *On understanding grammar.* Academic Press, New York.

Givón, T. 2002. *Bio-linguistics.* John Benjamins, Amsterdam.

Goh, K.-I., E.E. Oh, H. Jeong, B. Kahng, and D. Kim. 2002. Classification of scale-free networks. *Proc. Nat. Acad. Sci. USA*, 99:12583.

Gold, E. M. 1967. Language identification in the limit. *Information and Control*, 10:447–474.

Goldberg, Adele. 1995. *Constructions: a construction grammar approach to argumetn structure.* Chicago University Press, Chicago.

Gould, S. J. 1987. The limits of adaptation: is language a spandrel of the human brain? paper presented to the Cognitive Science Seminar, Center for Cognitive Science, MIT.

Gould, S. J. and N. Eldredge. 1993. Punctuated equilibrium comes of age. *Nature*, 336:223–227.

Grassly, Nicholas C., Arndt von Haeseler, and David C. Krakauer. 2000. Error, population structure and the origin of diverse sign systems. *J. theor. Biol.*, 206:369–378.

Greenberg, J. 1968. Some universals of grammar. In J. Greenberg, editor, *Universals of language.* MIT Press, Cambridge.

Greenberg, J. H. 1966. *Language Universals: with Special Reference to Feature Hierarchies.* Mouton.

Greenfield, P. M. and E. S. Savage-Rumbaugh. 1991. Imitation, grammatical development and the invention of protogrammar. In N. Krasnegor amd D. M. Rumbaugh, M. Studdert-Kennedy, and R.L. Scheifelbusch, editors, *Biological and behavioral determinants of language development.* Erlbaum, Hillsdale, New Jersey, pages 235–258.

Haken, H. 1979. *Synergetics-an introduction: nonequilibrium phase transitions & self-Organization in physics, chemistry & biology.* Springer-Verlag, New York.

Hall, R. A. 1953. Haitian creole: Grammar, texts, vocabulary. *American Folcklore Society Memoire*, (43).

Harnad, S. 1990. The symbol grounding problem. *Physica D*, 42:335–346.

Harremoës, P. and F. Topsøe. 2001. Maximum entropy fundamentals. *Entropy*, 3:227–292.

Harremoës, P. and F. Topsøe. 2002. Zipf's law, hyperbolic distributions and entropy loss. In *IEEE International Symposium on Information Theory*. in press.

Hauser, Marc D. 1996. *The evolution of communication*. MIT Press, Cambridge, MA.

Hauser, Marc D., Noam Chomsky, and W. Temcuseh Fitch. 2002. The faculty of language: what is, who has it and how did it evolve? *Science*, 298:1569–1579.

Hauser, Marc D. and Douglas A. Nelson. 1991. Intentional signaling in animal communication. *TREE*, 6(6):186–189.

Hawkins, J. A. 1994. *A performance theory of order and constituency*. Cambridge University Press, New York.

Hawkins, John A. 1992. On the evolution of human language. In John A. Hawkins and Murray Gell-Mann, editors, *Innateness and function in language universals*, pages 87–120, Redwood, CA. Addison Wesley.

Hays, D. 1964. Dependency theory: a formalism and some observations. *Language*, 40:511–525.

Helbig, G. 1992. *Probleme der Valenz und Kasustheorie*. Niemeyer, Tübinguen.

Herman, L. M., D. G. Richards, and J. P. Wolz. 1984. Comprehension of sentences by bottlenosed dolphins. *Cognition*, 16:129–219.

Hudson, Richard. 1984. *Word Grammar*. Blackwell, Oxford.

Hudson, Richard. 1990. *English word grammar*. Blackwell, Oxford.

Hudson, Richard. 1999a. *Envyclopedia of English grammar and word grammar*. http://www.phon.ucl.ac.uk/home/dick/papers.htm.

Hudson, Richard. 1999b. Review of Terrence Deacon 'The symbolic species: the co-evolution of language and the human brain'. London: Penguin, 19 *Journal of Pragmatics*, August.

Hurdord, Jim. 2002. Expression/induction models of language evolution: dimensions and issues. In Ted Briscoe, editor, *Linguistic evolution through language acquisition*. Cambridge University Press, Cambridge, pages 301–344.

Hurford, J. R., M. Studdert-Kennedy, and C. Knight, editors. 1998. *Approaches to the evolution of language: social and cognitive bases*. Cambridge University Press, Cambridge.

Hurford, Jim. 1998. Review of 'the symbolic species: the co-evolution of language and the human brain', by terrence deacon (1997, penguin press). *The Times Literary Supplement*, October.

Hřebíček, Luděk. 1995. Text levels. language constructs, constituents and the Menzerath-Altmann law. *Quantitative Linguistics*, 56.

Jackendoff, R. 1994. *Patterns in the mind*. Basic Books.

Jackendoff, Ray. 2002. *Foundations of language*. Oxford University Press.

Jacob, F. 1977. Evolution and tinkering. *Science*, 196:1161–1166.

Janik, V. M. 2000. Food-related bray calls in wild bottlenose dolphins (*tursiops truncatus*). *Proc. R. Soc. Lond. B.*, 267:923–927.

Janik, Vincent M. 1999. Pitfalls in the categorization of behavior: a comparison of dolphin whistle classification methors. *Anim. Behav.*, 57:113–143.

Janik, Vincent M. and Peter J. B. Slater. 1998. Context-specific use suggests that bottlenose dolphin signature whistless are cohesion calls. *Anim. Behav.*, 56:829–838.

Jenkins, Lyle. 2000. *Biolinguistics. Exploring the biology of language*. Cambridge University Press, Cambridge, UK.

Jeong, H., S.P. Mason, A.-L. Barabási, and Z. N. Oltvai. 2001. Lethality and centrality in protein networks. *Nature*, 411:41–42.

Jeong, H., B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. 2000. The large-scale organization of metabolic networks. *Nature*, 407:651–654, October.

Johnson, Carolyn, Henry Davis, and Marlys Macken. 1999. Symbols and structures in language acquisition. In Andrew Lock and Charles R. Peters, editors, *Handbook of human symbolic evolution*. Blackwell, Oxford, pages 686–746.

Kanter, I. and D. A. Kessler. 1995. Markov processes: linguistics and zipf's law. *Phys. Rev. Lett.*, 74:4559–4562.

Kapur, J. N., 1989a. *Maximum entropy models in science and engineering*, chapter Maximum-entropy discrete univariate probability distributions, pages 30–43. Wiley, New Delhi.

Kapur, J. N. 1989b. *Maximum entropy models in science and engineering*. Wiley, New Delhi.

Kauffman, S. A. 1993. *The Origins of Order: Self-Organization*. Oxford University Press, New York.

Kay, Paul and Charles Fillmore. 1990. Grammatical constructions and linguistic generalizations: the what's x doing y construction. *Language*, 75:1–33.

Kayne, Richard S. 1994. *The Antisymmetry of Syntax*. The MIT Press, Cambridge.

King, Barbara J. 1994. *The information continuum: evolution of social information transfer in monkeys, apes and hominids.* School of American Research Press, Santa Fe, NM.

Kinouchi, O., A. S. Martinez, G. F. Lima, G. M. Lourenço, and S. Risau-Gusman. 2002. Deterministic walks in random networks: an application to thesaurus graphs. *Physica A*, 315:665–676.

Kirby, S. and J. Hurford. 2001. The emergence of linguistic structure: An overview of the iterated learning model. In Angelo Cangelosi and Domenico Parisi, editors, *Simulating the Evolution of Language.* Springer Verlag, London, pages 121–148.

Kirby, Simon. 2000. Syntax without natural selection: how compositionality emerges from vocabulary in a population of learners. In C.Knight *et al.*, editor, *The evolutionary emergence of language.* Cambridge University Press, Cambridge, pages 303–322.

Kirby, Simon. 2001. Spontaneous evolution of linguistic structure. an iterated learning model of the emergence of regularity and irregulariy. *IEEE Trans. Evol. Comp.*, 5(2).

Kirby, Simon. 2002a. Learning, bottlenecks and the evolution of recursive syntax. In Ted Briscoe, editor, *Linguistic evolution through language acquisition.* Cambridge University Press, Cambridge, pages 173–204.

Kirby, Simon. 2002b. Natural language from artificial life. *Artificial Life*, 8(2):185–215.

Kirby, Simon. 2002c. The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In Ted Briscoe, editor, *Linguistic evolution through language acquisition.* Cambridge University Press, Cambridge, pages 111–172.

Köhler, Reinhard. 1986. *Zur Linguistischen Synergetik: Struktur und Dynamik der Lexik.* Brockmeyer, Bochum.

Köhler, Reinhard. 1987. System theoretical linguistics. *Theor. Linguist.*, 14(2-3):241–257.

Köhler, Reinhard. 1999. Syntactic structures: properties and interrelations. *J. Quantitative Linguistics*, 6:46–57.

Köhler, Reinhard and Gabriel Altmann. 2000. Probability distributions of syntactic units and properties. *J. Quantitative Linguistics*, 7:189–200.

Komarova, Natalia and Martin A. Nowak. 2001. The evolutionary dynamics of the lexical matrix. *Bulletin of Mathematical Biology*, 63:451–484.

Koren, Y. and D. Harel. 2002. A multi-scale algorithm for the linear arrangement problem. In *Proceedings of 28th Inter. Workshop on Graph-Theoretic Concepts in Computer Science (WG'02)*, volume 2573 of *Lecture Notes in Computer Science*, pages 293–306. Springer Verlag.

Krakauer, David C. 2001. Selective imitation for a private sign system. *J. theor. Biol.*, 213:145–157.

Langacker, Ronald. 1987. *Foundations of cognitive grammar 1: theoretical prerequisites.* Stanford University Press, Stanford.

Langacker, Ronald. 1990. *Concept, image and symbol. The cognitive basis of language.* Mouton the Gruytier, Berlin.

Lepold, Edda. 1998. Frequency spectra within word-length classes. *J. Quantitative Linguistics*, 5(3):224–231.

Li, Charles N., editor. 1976. *Subject and Topic.* Academic PRess, New York.

Li, W. 1998. Letters to the editor. *Complexity*, 3:9–10. Comments to "Zipf's Law and the structure and evolution of languages" A.A. Tsonis, C. Schultz, P.A. Tsonis, COMPLEXITY, 2(5). 12-13 (1997).

Li, Wentian. 1992. Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE T. Inform. Theory*, 38(6):1842–1845, November.

Lieberman, P. 1991a. *Uniquely Human: The evolution of speech, thought and selfless behavior.* Harvard University Press, Cambridge, MA.

Lieberman, P. 1991b. *Uniquely Human: The evolution of speech, thought and selfless behavior.* Harvard University Press., Cambridge, MA.

Lieberman, Philip. 1992. On the evolution of human language. In John A. Hawkins and Murray Gell-Mann, editors, *The Evolution of Human Languages*, pages 21–47, Redwood, CA. Addison Wesley.

Lieberman, Philip and Stephen M. Kosslyn. 2002. *Human Language and Our Reptilian Brain: The Subcortical Bases of Speech, Syntax, and Thought.* Perspectives in Cognitive Neuroscience. Harvard University Press, Cambridge, MA.

Losick, R. and D. Kaiser. 1997. Why and how bacteria communicate. *Sci. Am.*, pages 68–73.

Mandelbrot, B. 1953. An informational theory of the statistical structure of language. In W. Jackson, editor, *Communication theory.* Butterworths, London, page 486.

Mandelbrot, B. 1966. Information theory and psycholinguistics: A theory of word frequencies. In P. F. Lazarsfield and N. W. Henry, editors, *Readings in mathematical social sciences.* MIT Press, Cambridge, pages 151–168.

Manning, Christopher D. and Hinrich Schütze, 1999. *Foundations of statistical natural language processing*, chapter Introduction. MIT Press, Cambridge, MA.

Mathias, Nisha and Venkatesh Gopal. 2001. Small worlds: How and why. *Phys. Rev. E*, 63:021117–021128.

Ferrer i Cancho, R., C. Janssen, and R. V. Solé. 2001. Topology of technology graphs: small world patterns in electronic circuits. *Phys. Rev. E*, 64:046119–046124.

Ferrer i Cancho, R. and Ricard V. Solé. 2003. Least effort and the origins of scaling in human language. *Proc. Natl. Acad. Sci. USA*, 100:788–791.

Ferrer i Cancho, Ramon. 2003. Decoding least effort and scaling in signal frequency distributions. In preparation.

Ferrer i Cancho, Ramon, Béla Bollobás, and Oliver Riordan. 2003. Zipf's law consequences for syntax and symbolic reference. Submitted to Biology Letters.

Ferrer i Cancho, Ramon and Ricard V. Solé. 2001. Two regimes in the frequency of words and the origin of complex lexicons: Zipf's law revisited. *J. Quantitative Linguistics*, 8(3):165–173.

Ferrer i Cancho, Ramon and Ricard V. Solé, 2003. *Optimization in complex networks*. Springer, Berlin.

Ferrer i Cancho, Ramon, Ricard V. Solé, and Reinhard Köhler. 2003. Patterns in syntactic dependency networks. Submitted to Phsysical Review E.

Ferrer i Cancho, Ramon and Ricard V. Solé. 2001. The small-world of human language. *Proc. R. Soc. Lond. B*, 268:2261–2266.

McCowan, Brenda, Laurence R. Doyle, and Sean F. Hanser. 2002. Using information theory to assess the diversity, complexity and development of communicative repertoires. *Journal of Comparative Psychology*, 116:166–172.

McCowan, Brenda, Sean F. Hanser, and Laurance R. Doyle. 1999. Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires. *Anim. Behav.*, 57:409–419.

Medawar, P. B. 1969. *Induction and Intuition in Scientific Thought*. Methuen and Co., London.

Meinhardt, Hans. 1995. *The algorithmic beauty of see shells*. Springer-Verlag, Berlin.

Melčuck, Igor. To appear. Dependency in linguistic description.

Melčuk, Igor. 1988. *Dependency Syntax: Theory and Practice*. SUNY.

Melčuk, Igor. 1989. *Dependency grammar: theory and practice.* New York, University of New York.

Melčuk, Igor. 2002. Language: Dependency. In N. J. Smelser and Paul B. Baltes, editors, *International Encyclope-dia of the Social and Behavioral Sciences.* Pergamon, Oxford, pages 8336–8344.

Michell, John F. 1999. *Who wrote Shakespeare?* Thames & Hudson, Slovenia.

Miller, G. 1981. *Language and speech.* W. H. Freeman and Co., San Francisco.

Miller, George. A. 1957. Some effects of intermittent silence. *Am. J. Psychol.*, 70:311–314.

Miller, George A. and Noam Chomsky. 1963. Finitary models of language users. In R. D. Luce, R. Bush, and E. Galanter, editors, *Handbook of Mathematical Psychology*, volume 2. Wiley, New York.

Miller, George A. and Patricia M. Gildea. 1987. How children learn words. *Scientific American*, 257(3):94–99.

Miller, R. V. 1998. Bacterial gene swapping in nature. *Sci. Am.*, 278.

Mitchinson, G. 1991. Neural branching patterns and the economy of cortical wiring. *Proc. R. Soc. London B*, 245:151–158.

Montemurro, Marcelo A. 2001. Beyond the Zipf-Mandelbrot law in quantitative linguistics. *Physica A*, 300:567–578. cond-mat/0104066.

Montroll, E. W. and M. F. Shlesinger. 1983. Maximum entropy formalism, fractals, scaling phenomena, and $1/f$ noise: a tale of tails. *J. Stat. Phys.*, 32:209–230.

Morowitz, H. J., J. D. Kostelnik, J. Yang, and G. D. Cody. 2000. The origin of intermediary metabolism. *Proc. Natl. Acad. Sci. Sci. USA*, 97:7704–7708.

Motter, Adilson E., Alessandro P. S. de Moura, Ying-Cheng Lai, and Partha Dasgupta. 2002. Topology of the conceptual network of language. *Phys. Rev. E*, 65:065102.

Murray, J. D. 1980. *Mathematical biology.* Springer-Verlag, Berlin.

Naranan, S. 1992. Statistical laws in information science, language and system of natural numbers: some striking similarities. *Journal of Scientific and Industrial Research*, 51:736–755.

Naranan, S. and V. K. Balasubrahmanyan. 1993. Information theoretic model for frequency distribution of words and speech sounds (phonemes) in language. *Journal of Scientific and Industrial Research*, 52:728–738.

Naranan, S. and V. K. Balasubrahmanyan. 2000. Information theory and algorithmic complexity: applications to linguistic discourses and DNA sequences as complex systems. part i: Efficiency of the genetic code of DNA. *J. Quantitative Linguistics*, 7(2):129–151.

Naranan, S. and V.K. Balasubrahmanyan. 1992a. Information theoretic models in statistical linguistics - Part I: A model for word frequencies. *Current Science*, 63:261–269.

Naranan, S. and V.K. Balasubrahmanyan. 1992b. Information theoretic models in statistical linguistics - part ii: Word frequencies and hierarchical structure in language. *Current Science*, 63:297–306.

Naranan, S. and V.K. Balasubrahmanyan. 1998. Models for power law relations in linguistics and information science. *J. Quantitative Linguistics*, 5(1-2):35–61.

Newman, M. E. J. 2000. Models of the small-world. *Journal of Statistical Physics*, 101(3/4):819–841.

Newman, M. E. J. 2001. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci.*, 98:404–409.

Newman, M. E. J. 2002. Assortative mixing in networks. *Phys. Rev. Lett.*, 89:208701.

Newman, M. E. J. 2003a. Mixing patterns in networks. *Phys. Rev. E 67*, page 026126.

Newman, M. E. J. 2003b. The structure and function of complex networks. *SIAM Review*, pages 167–256.

Newman, M. E. J., S. H. Strogatz, and D. J. Watts. 2001. Random graphs with arbitrary degree distribution and their applications. *Phys. Rev. E*, 64:026118.

Newmeyer, F. 1991. Functional explanations in linguistics and the origins of language. *Lang. Commun.*, 11:3–96.

Newmeyer, F. J. 2000. Reconstructing "proto-world" word order. In C.Knight *et al.*, editor, *The evolutionary emergence of language*. Cambridge University Press, Cambridge, pages 372–388.

Newport, E. L. 1990. Maturational constraints on language learning. *Cognitive Science*, 14:11–28.

Nicolis, J. S. 1991. *Chaos and information processing*. World Scientific, Singapore.

Nishikawa, Takashi, Adilson E. Motter, Ying-Cheng Lai, and Frank C. Hoppensteadt. 2002. Smallest small-world network. *Physical Review E*, 66:046139.

Noad, Michael, Douglas H. Cato, M.M. Bryden, Micheline-N. Jenner, and K. Curt S. Jenner. 2000. Cultural revolution in whale songs. *Nature*, 408:537.

Nowak, M. A., Komarova, and P. Niyogi. 2002. Computational and evolutionary aspects of language. *Nature*, 417:611–617.

Nowak, M. A. and N. L. Komarova. 2001. Towards an evolutionary theory of language. *Trends in Cognitive Sciences*, 5(7):288–295.

Nowak, Martin A. 2000a. The basic the reproductive ratio of a word, the maximum the size of the lexicon. *J. theor. Biol.*, 204:179–189. doi:10.1006/jtbi.2000.1085.

Nowak, Martin A. 2000b. Evolutionary biology of language. *Phil. Trans. R. Soc. Lond. B*, 355:1615–1622.

Nowak, Martin A., Natalia L. Komarova, and Partha Niyogi. 2001. Evolution of universal grammar. *Science*, 291:114–118.

Nowak, Martin A. and David C. Krakauer. 1999. The evolution of language. *Proc. Natl. Acad. Sci. USA*, 96:8028–8033, July.

Nowak, Martin A., David C. Krakauer, and Andreas Dress. 1999. An error limit for the evolution of language. *Proc. R. Soc. London B*, 266:2131–2136.

Nowak, Martin A., J. B. Plotkin, and V. A. Jansen. 2000. The evolution of syntactic communication. *Nature*, 404:495–498.

Nowak, Martin A., Joshua B. Plotkin, and David C. Krakauer. 1999. The evolutionary language game. *J. theor. Biol.*, 200:147–162.

Oliphant, Michael. 2002. Learned systems of arbitrary reference: the foundation of human linguistic uniqueness. In Ted Briscoe, editor, *Linguistic evolution through language acquisition*. Cambridge University Press, Cambridge, pages 23–52.

Peirce, C. 1932. *Collected papers of Charles Sanders Peirce. Volume 2: Elements of logic*. Harvard University Press, Cambridge, MA.

Pietronero, L., E. Tosatti, V. Tosatti, and A. Vespignani. 2001. Explaining the uneven distribution of number in nature: the laws of Benford and Zipf. *Physica A*, 293:297–304.

Pinker, S. 1996. *The language instinct*. HarperCollins, New York.

Pinker, S. and P. Bloom. 1990. Natural language and natural selection. *Behav. Brain Sci.*, 13:707–784.

Poeppel, David. 1997. Mind over chatter. *Nature*, 388:734.

Polya, G. 1931. Sur quelques points de la théorie des probabilités. *Ann. Inst. H. Poincaré.*

Popper, K. R. 1968. *The Logic of Scientific Discovery.* Harper and Row, New York.

Pritchet, B. 1992. *Grammatical competence and parsing performance.* The University of Chicago, Chicago.

Prün, Claudia. 1999. G. K. Zipf's conception of language as an early prototype of synergetic linguistics. *J.Quantitative Linguistics*, 6:78–84.

Pulvermuller, F. 1999. Words in the brain's language. *Behavioral and Brain Sciences*, 22:253–336.

Pulvermuller, F. 2001. Brain reflections of words and their meaning. *Trends in Cognitive Sciences*, 5(12):517–524.

Ramsden, J. J. and J. Vohradský. 1998. Zipf-like behavior in procariotic protein expression. *Physical Review E*, 58:7777–7780.

Rapoport, A. 1982. Zipf's law re-visited. *Quantitative Linguistics*, 16:1–28.

Ravasz, E. and A.-L. Barabási. 2002. Hierarquical organization in complex networks. *Phys. Rev. E*, 67:026112.

Ravasz, E., A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási. 2002. Hierarquical organization of modularity in metabolic networks. *Science*, 297:1551–1555.

Ravin, Yael and Claudia Leacock. 2000a. Polysemy: an overview. In Yael Ravin and Claudia Leacock, editors, *Polysemy. Theoretical and computational approaches.* Oxford University Press, New York.

Ravin, Yael and Claudia Leacock, editors. 2000b. *Polysemy. Theoretical and computational approaches.* Oxford University Press, New York.

Reader, S. M. and K. N. Laland. 2002. Social intelligence, innovation and enhanced brain size in primates. *Proc. Nat. Acad. Sci. USA*, 99:4436–4441.

Reder, L.M., J. R. Anderson, and R. A. Bjork. 1974. A semantic interpretation of encoding specificity. *Journal of Experimental Psychology*, 102:648–656.

Reich, Peter A. 1986. *Language development.* Prentice-Hall, Englewood Cliffs, NIJ.

Riedemann, Hagen. 1996. Word-length distribution in english press texts. *J. of Quantitative Linguistics*, 3(3):265–271.

Rodriguez-Iturbe, I. and A. Rinaldo. 1997. *Fractal River Basins.* Cambridge U. Press, Cambridge.

Romaine, S. 1988. *Pidgin and Creole Languages.* Longman, London.

Romaine, S. 1992. The evolution of linguistic complexity in pidgin and creole languages. In John A. Hawkins and Murray Gell-Mann, editors, *The Evolution of Human Languages*, pages 213–238. Addison Wesley.

Savage-Rumbaugh, Sue. 1999. Ape language. between a rock and a had place. In Barbara J. King, editor, *The origins of language. What Nonhuman primates can tell us.* School of American Research Press, Santa Fe, NM, pages 115–188.

Schuster, P. 2001. Taming combinatorial explosion. *Proc. Natl. Acad. Sci. Sci. USA*, 97:7678–7680.

Seyfarth, R., D. Cheney, and P. Marler. 1980a. Monkey responses to three different alarm calls: evidence of predator classification and semantic communication. *Science*, 210:801–803.

Seyfarth, Robert M., Dorotey Cheney, and Peter Marler. 1980b. Vervet monkey alarm calls: semantic communication in a free-ranging primate. *Anim. Behav.*, 28:1070–194.

Shannon, Claude E. 1948. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423 623–656.

Shiloach, Yossi. 1979. A minimum linear arrangement algorithm for undirected trees. *SIAM J. Comput.*, 8(1):15–32.

Sigman, M. and G. A. Cecchi. 2002. Global organization of the wordnet lexicon. *Proc. Natl. Acad. Sci. USA*, 99(3):1742–1747.

Simon, Herbert A. 1955. On a class of skew distribution functions. *Biometrika*, 42:425–440.

Sinha, Christopher G. 1999. Theories of simbolization and devolopment. In Andrew Lock and Charles R. Peters, editors, *Handbook of human symbolic evolution.* Blackwell, Oxford, pages 483–500.

Sipser, Michael. 1999. *Introduction to the theory of computation.* PWS Publishing Company, Boston, MA.

Sleator, D. and D. Temperley. 1993. Parsing english with a link grammar. In *Third international workshop on parsing technologies.*

Sleator, Daniel and Davy Temperley. 1991. Parsing English with a link grammar. Technical report, Carnegie Mellon University.

Smith, K., S. Kirby, and H. Brighton. To appear. Iterated learning: a framework for the emergence of language. In C. Hemelrijk, editor, *Self-organization and Evolution of Social Behaviour.* Cambridge University Press, Cambridge, UK.

Sokal, Robert R. and F. James Rohlf. 1995. *Biometry. The principles and practice of statistics in biological research.* W. H. Freeman and Co., New York.

Solé, R. V. and O. Miramontes. 1995. Information at the edge of chaos in fluid neural networks. *Physica D*, 80:171–180.

Solé, R. V., R. Pastor-Satorras, E. Smith, and T. Kepler. 2002. A model of large-scale proteome evolution. *Adv. Complex Syst.*, 5:43–54.

Solé, Ricard V., Susanna C. Manrubia, Bartolo Luque, Jordi Delgado, and Jordi Bascompte. 1996. Phase transitions and complex systems. *Complexity*, 1(4):13–26.

Stanley, H. E., L.A.N. Amaral, S.V. Buldyrev, A.L. Goldberger, S. Havlin, H. Leschhorn, P. Maas, H. A. Makse, C.-K. Peng, M.A. Salinger, M. H. R. Stanley, and G. M. Viswanathan. 1996. Scaling and universality in animate and inanimate systems. *Physica A*, 231:20–48.

Stanley, H.E., L.A.N. Amaral, P. Gopikrishnan, P. Ch. Ivanov, T. H. Keitt, and V. Plerou. 2000. Scale invariance and universality: organinzing principles in complex systems. *Physica A*, 281:60–68.

Stauffer, Dietrich and A. Aharony. 1994. *Introduction to Percolation Theory.* Taylor & Francis, London.

Steele, Susan. 1978. Word order variation, a typological study. In Joseph H. Greenberg, Charles A. Ferguson, and Edith A. Moravcsik, editors, *Universals of language: Syntax*, volume 4. Stanford University Press, Stanford, CA, pages 585–623.

Steels, L. 1996. Self-organizing vocabularies. In C. Langton, editor, *Proceedings of Alife V*, Nara Japan.

Steyvers, M. and J. Tenenbaum. 2001. The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *cond-mat/0110012.*

Strogatz, Steven H. 2001. Exploring complex networks. *Nature*, 410:268–276.

Suzuki, Ryuji, Peter L. Tyack, and John Buck. 2003. The use of Zipf's law in animal communication analysis. *Anim. Behav.* Accepted.

Tsonis, A. A., C. Schultz, and P. A. Tsonis. 1997. Zipf's law and the structure and evolution of language. *Complexity*, 3(5).

Tuldava, J. 1996. The frequency spectrum of text and vocabulary. *J. Quantitative Linguistics*, 3(1):38–50.

Těšitelová, M. 1985. Kvantitativní charakteristiky současné češtiny (quantitative characteristics of present-day czech). *Academia Praha*, page 249s.

Ueno, M. and M. Polinsky. 2002. Maximizing processing in an SOV language: A corpus study of Japanese and English. In *15th annual CUNY conference on human sentence processing*, New York, March. CUNY.

Uhlírova, L., I. Nebeská, and J. Králík. 1982. Computational data analysis for syntax. In J. Horecký, editor, *COLING 82, Proceedings of the Ninth International Conference on Computational Linguistics. Prague July 5-10.* North-Holland Publishing Company, Amsterdam, pages 391–396.

Ujhelyi, Maria. 1996. Is there any intermediate stage between animal commnication and language? *J. theor. Biol.*, 180:71–76.

Uriagereka, Juan. 1998. *Rhyme and Reason. An introduction to Minimalist Syntax.* The MIT Press, Cambridge, Massachusetts.

Valverde, Sergi, Ramon Ferrer i Cancho, and Ricard V. Solé. 2002. Scale free networks from optimal design. *Europhysics Letters*, 60:512–517.

Vázquez, A., A. Flammini, A. Maritan, and A. Vespignani. 2003. Modeling of protein interaction networks. *Complexus*, 1:38–44.

Vihman, Marilyn M. and Rory A. Depaolis. 2000. The role of mimesis in infant language development: evidence for phylogeny? In C.Knight *et al.*, editor, *The evolutionary emergence of language.* Cambridge University Press, Cambridge, pages 130–145.

von Humboldt, W. 1972. *Linguistic variability and intellecual development.* Univ. of Penssylvania Press, Philadelphia.

Watts, Duncan J. and Steven H. Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, June.

Werker, Janet F. and Athena Vouloumanos. 2001. Speech and language processing in infancy: a neurocognitive approach. In Charles A. Nelson and Monica Luciana, editors, *Handbook of developmental cognitive neuroscience.* MIT Press, Cambridge, MA.

West, G. B., J. H. Brown, and B. J. Enquist. 1997. A general model for the origin of allometric scaling laws in biology. *Science*, 276(107):122–126.

Wimmer, G. and G. Altmann. 1996. The theory of word length: some results and generalizations. *Glottometrika*, 15:112–133.

Wimmer, G., R. Köhler, R. Grotjahn, and G. Altmann. 1994. Towards a theory of word length distribution. *J. Quantitative Linguistics*, 1:98–106.

Wimmer, Gejza and Gabriel Altmann. 1999. *Thesaurus of univariate discrete probability distributions.* STAMM Verlag, Germany.

Wolfram, Stephen. 2002. *A new kind of science.* Wolfram Media, Champaign.

Zanette, Damián H. and Susanna C. Manrubia. 2001. Vertical transmission of culture and the distribution of familiy names. *Physica A*, 295(1-2):1–8.

Zipf, G. K. 1932. *Selected studies of the principle of relative frequency in language.* Hardvard University Press, Cambridge (Mass.).

Zipf, G. K. 1935. *The psycho-biology of language.* Houghton Mifflin, Boston.

Zipf, G. K. 1972a. *Human behaviour and the principle of least effort. An introduction to human ecology.* Hafner reprint, New York. 1st edition: Cambridge, MA: Addison-Wesley, 1949.

Zipf, G. K. 1972b. *Human behaviour and the principle of least effort. An introduction to human ecology.* Hafner reprint, New York. 1st edition: Cambridge, MA: Addison-Wesley, 1949.

Zipf, George Kingsley. 1942. Children's speech. *Science*, 96:344–345.

Zörnig, Peter and Gabriel Altmann. 1995. Unified representations of zipf distributions. *Computational Statistics and Data Analysis*, 19:461–473.