A unified approach to the emergence of complex communication

Bernat Corominas Murtra

TESI DOCTORAL UPF / ANY 2011i

DIRECTOR DE LA TESI
Ricard Solé. Departament Departament de Ciències Experimentals i de la Salut

...en aquell temps (...) no corria tanta pedanteria com ara i la gent no tractava de dissimular amb tesis, missatges ni teories abstractes el fons apassionat que tots portem dintre.

Joan Sales

A la Isabel

Agraïments

He tingut sort de viure en una època on he pogut estudiar, he tingut la sort d'aprendre l'amor pel rigor i el saber pausat dels meus pares, l'Agustí i la Rosa, que em deien (en grec) que "les coses belles són difícils". He tigut la sort de poder compartir-ho (quasi bé) tot amb l'Andreu. He tingut més sort que els meus avis, Frederic, Paquita, Ramon i Montserrat. El Frederic va voler estudiar física, la Montserrat, matemàtiques, però que les absurditats de la història els hi ho van impedir. Espero que aquesta tesi sigui digne de tot el que m'han donat.

He tingut la sort de compartir la quotidianitat amb l'Agnès, l'Eugeni, el Miquel Àngel, la Maria, l'Anna, la Núria, la Laura, l'Oriol, el Lluc i l'Alfons Bertran. Ha sigut meravellós compartir ciència -i d'altres coses- amb el Josep, el Carlos, l'Andreea, el Sergi, el Javier, el Jordi Fortuny, la Stefanie, el Jordi Delgado, el Joaquín, el Harold, l'Steen, el Kepa, el Martí Sànchez el Martí Rosas, el Ben, el Salva, la Núria, la Mireia, el Txuss, el Jeroni i la Iraia. La sort de comptar amb un equip de suport a tots nivells en el que hi ha l'Alfons Gonzàlez, el Miquel, la Carina, la Rosa, la Natàlia, el Carles, la Sussana, la Sònia i l'Ainara.

Perquè una tesi és un garbuix -més desordenat del què acceptariem- d'esforços, casualitats, il.lusions i desenganys. Però per sobre de tot, acabar una tesi és questió de sort. Em vaig sentir afortunat quan el

Ricard em cridà al seu costat. Les versions del moment són contradictòries: ell defensa que em va trobar convertit en un personatge de Txèkhov a la porta d'un Teatre de províncies. Jo defenso que buscava personatges per portar al teatre una obra de Dostoievski, i que vaig ser jo el que el vaig triar gràcies a la seva -indiscutible- indoneïtat. Al final, ambdues són veritat, perquè ambdues són inventades, i sempre he professat respecte per les històries inventades -invents que alguns anomenen eufemísticament "literatura", o "ciència". Perquè, no ens enganyem, el motor del científic -com el de l'escriptor, o el músic- és l'experiència de la llibertat de crear i inventar.

No és una invenció dir que sense la Isabel no hauria fet aquesta tesi. No és un exageració, tampoc: és un fet. Ella m'acompanya per un camí que espero que sigui molt llarg. Diria que Martí i el Biel no s'han interessat per la meva tesi. Ells posseeixen la intel.ligència verdadera, la que no s'amaga darrera abstraccions, la directa, la instintiva. Ells m'estimen, i és que, hi ha mostra més sublim d'intel.ligència que l'amor? no ho crec.

Abstract

This dissertation studies the emergence of complexity in natural codes taking human language as the object of study. We focus our analysis in i) Statistical patterns of complexity, ii) Generative mechanisms and iii) Information theoretic aspects of complex communication. We first provide a quantitative identification of the emergence of syntax at the ontogenetic level through modern theory of complex networks. We then propose a mathematical backbone for human syntax with the aim to identify the minimal formal properties for a natural generative system, which is consistent with the previous observed patterns. We follow by studying a well-known statistical pattern of complex communication systems, Zipf's law, for which we propose an information-theoretic argument accounting for its emergence. Finally, the problem of referentiality is studied, proposing an information-theoretic functional to evaluate its degree of conservation in a given communicative exchange.

Resum

Aquesta tesi estudia l'emergència de complexitat en sistemes de comunicació naturals, prenent el llenguatge humà com a principal objecte d'estudi. Ens centrem en i) Patrons estadístics de complexitat, ii) Mecanismes generatius i iii) aspectes relacionats amb la teoria de la informació. Primer mostrem un estudi on es quantifica l'emergència de la sintaxi al nivell ontogenètic usant la moderna teoria de xarxes complexes. Posteriorment, es proposa un esquelet matemàtic per a la sintaxi humana amb el propòsit d'identificar les mínimes propietats formals d'un sistema generatiu, essent aquest constructe consistent amb els patrons observats prèviament. Seguidament explorem un patró molt comú en sistemes de comunicació complexes, la llei de Zipf, presentant un argument que explica la seva emergència des de consideracions únicament basades en la teoria de la informació. Finalment, abordem el problema de la referencialitat, proposant un funcional consistent amb la teoria de la informació que evalua el seu grau de conservació en un intercanvi comunicatiu arbitrari.

Contents

Figure Index xvii

1 INTRODUCTION: FOUR KEY QUESTIONS 3
1.1 Life, Information and language 3
1.2 Language: A privileged window 5
1.2.1 Setting up the problem: Four questions 6
1.3 Generative mechanisms and patterns of unboundedness 9
1.3.1 Language acquisition 10
1.3.2 The minimal generative system 14
1.3.3 Zipf’s law 17
1.4 The Problem of Referentality 21

2 THEORETICAL FRAMEWORK 25
2.1 Entropy and information measures 27
2.1.1 Shannon Entropy 28
2.1.2 Kullback-Leibler Divergence 30
2.1.3 Conditional and Joint Entropies 32
2.1.4 Mutual Information 33
2.1.5 Channel Capacity 35
2.1.6 Channel Capacity and generative codes 37
2.2 Computational Complexity 42
2.2.1 Kolmogorov Complexity 43
2.2.2 Defining complexity 46

3 RESULTS AND DISCUSSION 51
3.1 The Ontogeny of Syntax 54
3.1.1 Patterns 55
3.1.2 Origins of the dynamical pattern 57
3.2 Generative mechanisms: Merge 58
3.2.1 Nesting grammars 60
3.2.2 Merge and the Acquisition process 62
3.2.3 A general framework 66
3.3 Zipf’s Law = unbounded complexity 66
3.3.1 Zipf’s Law as a nonequlibrium attractor . . . 67
3.3.2 Zipf’s Law in the communicative context . . . 69
3.3.3 Zipf’s Law at the edge of Infinity 71
3.3.4 Consequences for open-ended evolution 73
3.4 The problem of referentiality 76
3.4.1 No self-consistency paradox 77
3.4.2 How to conserve referentiality 80

4 CONCLUSIONS 83

5 PAPERS ON LANGUAGE ACQUISITION 101

6 PAPERS ON THEORETICAL SYNTAX 167

7 PAPERS ON ZIPF’S LAW 195

8 PAPERS ON REFERENTIALITY 213

List of Figures

1.1 The brain experiences great changes within the early life of humans. Children brains have been studied (a) in searching for changing patterns of activity under given tasks. Here differences in speech processing cortical maps are depicted for the left hemisphere of three-month (top) and ten-month old infants. Here the color scale is related to the response activity in controlled responses to speech stimuli -from [56]. Differences have been reported across all childhood from different sources, although none seems to involve a dramatic change in network patterns correlated to syntax network architecture. Network patterns obtained from the aggregation of syntactic relations of child's utterances during different periods of language acquisition. The first network (b) belong to the so-called 2-words stage, and it is worth to note the tree-like structure of the net (except for a few crossings). The second network (c) is obtained after the so-called syntactic spurt, and its topology is both quantitatively and qualitatively different. Both in (b) and (c) we display the connected component as well as a zoomed subnetwork in order to display some of the specific items contained at the core of each graph. 12

1.2 Syntactic structures are nests, not chains of elements. In this picture we show how a nesting algorithm works using the metaphor of the physical recipient: In the example proposed in the main text, the first element to be merged is a small recipient embedded in a bigger one -which corresponds to the syntactic constituent $\langle \text{the}, \text{shoemaker} \rangle$. This operation can be hypothetically expanded in an unbounded way generating arbitrarily large nested structures. 15

1.3 George Zipf (right) first reported about the widespread presence of a special scaling law with a well-defined scaling exponent. This law can be observed in human language (b) and many other contexts. The plot displayed at left shows the frequency distribution of words from a written text in log-log scale. The linear trend in this diagram indicate the presence of a power law. 18

1.4 An example of the potential consequences of an absence of referential value. Two opposed events (detecting an individual of the same species or a predator) are coded in some way by $\mathbf{P}$ and deterministically decoded by $\mathbf{Q}$, shifting the referential values of the signals. In terms of classical information theory, this system could display maximum mutual information. However, if functionality of the communicative exchange is supposed, the selective consequences of this scenario -where friends and enemies can be mistaken- is completely disastrous. A ”perfectly wrong” identification of a predator as a friendly neighbor is a deadly one. . 22

2.1 Claude Shannon (left) the founding father of information theory, provided the mathematical basis for a general framework of information coding and decoding through channels. The theory provides a powerful approach to the problem in both natural and man-made systems. Other researchers, such as the russian genius Andréi Kolmogorov (right) also contributed to this area, allowing to connect complexity and computational complexity in mathematical terms. 26

2.2 A communication system. An external input (objects surrounding the agent or actions taken by other parts of the environment) is coded in some way in signals which, in turn, are sent through the channel. The decoder receives these signals, maybe with errors due to noise and further decodes them into an external output. The central square (dashed lines) represents the essential ingredients required to define information. This highlights the observation that no referential value is actually related to the definition of information (see text). 36

2.3 a) Let the big circle be the volume of the abstract space where points are possible signals. Red balls represent the region of the space occupied by the signal represented by the point located at the center: If some other signal is represented by a point inside the sphere, the probability of error is non-zero and these two signals can be confused with non-vanishing probability. Above center we show the space saturated of signals i.e., there is a distribution of points by which their associated spheres do not intersect. However, in the upper right corner, we observe that, above a give threshold in the number of signals, intersection between spheres is unavoidable, which can be informally related to the channel capacity. b) If, instead of increasing the number of signals we combine existing signals -which have a configuration where the error probability is zero- the space available for the signals is the cartesian product of two spaces identical to the original one, thus exponentially increasing the number of possible signals and linearly increasing the amount of information that can be carried out by the existing signals. 38

2.4 Algorithmic complexity $K(\mathbf{x})$ of a stochastic string $\mathbf{x}$, indicating a set of observations made on a system, is measured as the lenght in bits of the minimal program p required to reproduce such string. For a fair coin toss (a) which generates a completely random sequence, the computer Ψ (b) would run a program with a length equal to the lenght of the string (which is an upper bound here). Here the size of the alphabet Σ is two (i. e. $|\Sigma| = 2$) but an arbitrary sequence (c) $\mathbf{y}$ obtained from the successive observations over complex system would not be restricted by the binary description. Instead, a large range of n possible symbols would be used to define our string now. This is coded through a minimal program which, when applied to a computer (d), replicates the $n-$ary original sequence. The length of this minimal program, coded in bits, is the Kolmogorov Complexity of $\mathbf{y}$. 44

2.5 The problem of how to properly define complexity based on information theoretical approximations was addressed by physicists Murray Gell-mann (left) and Seth Lloyd, working together at the Santa Fe Institute. They defined the concept of effective complexity in terms of the length of a highly compressed description of its regularities. 48

3.1 Language structure and its evolution have been analyzed by many scholars over the centuries. Charles darwin compared languages with living species, which are able to change, split and get extinct. Such comparison, as shown by modern genomics, was much more accurate that Darwin himself would have suspected. Noam Chomsky (right) one of the fathers of modern linguistics, speculated about the presence of some ”hardwired” readiness for acquiring language that would have been shaped in our brains through natural selection. 52

3.2 Changes in the structure of syntax nets as obtained by using complex networks approaches. Here we show: the average path length (a) the number of words and links (b) the clustering (c) and (d) the final degree distribution after the transition. As shown in (a) and (c) a small world pattern emerges at the transition point. 56

3.3 A possible minimal chromatic configuration for two networks belonging (above) to the two words stage and (below) beyond the syntactic spurt. The so-called chromatic number of a net is the minimal number of different colors needed to paint the nodes in such a way that no node has any neighbor sharing the same color, and it is a very powerful indicator of network complexity. Interestingly, the two words stage defines a bipartite network where only two colors are needed to properly paint the net, suggesting that such a protogrammar is strongly constrained by compatibility relations. Beyond the two words stage, the chromatic number increases abruptly and the underlying grammar is by far more flexible, suggesting a qualitative, more than quantitative, shift on grammar's properties. 59

3.4 Stuart Kauffman (left) and Walter Fontana were among the first scientists pointing towards the problem of open-ended evolution in artificial systems and their relevance for evolution and evolvability. They also showed that complex evolved systems in-silico can display spontaneous emergence of hierachical organization. Such forms of organization can be formally approached in terms of formal languages. The artificially evolved systems (right) typically display feedbacks and multiple levels of dynamical structure. 74

Preface

This dissertation is mainly about the concept of ”infinite”. But not the Aristotle's ”absolute” or Cantor's ”alephs”, nor the ”limit of very large systems” of statistical mechanics. It tries to disentangle the features of another infinite: The one hypothetically created by life, which (maybe) built systems whose complexity is able to grow and grow unboundedly, the infinite of language, which is the infinite of Wilhem Von Humboldt and Noam Chomsky.

Obviously, from the scientific viewpoint, infinity will be always a working hypothesis or even a conjecture, since there is no way to test it empirically. But, what would be the footprints we could expect for a system which has this infinite potential? Could we define quantitative or formal indicators consistent to the hypothesis of infinite?

In the scientific inquiry, such concept has been named in different ways. In evolution, it has been designed by the name of ”open endedness”. In linguistics, ”recursion”. They both refer to the (hypothetical) ability of living systems or human language to generate an unbounded number of new structures. It is precisely human language that gives clues to the general problem of unboundedness, and this is the reason by which human language is at the center of this dissertation. However it is not a linguistics' dissertation. Instead, one of the ambitions of this dissertation is to show how exploring human language in a rigorous way we shed light to foundational problems of modern biology or, being more generic, to the science of complex systems.

The study of human language implies a departure from the common procedure of many fields of science, like physics. Indeed, although physics can be extremely helpful for the study of human language, there are, nowadays, fundamental differences between the philosophical attitude we adopt when doing research at the theoretical level. Indeed, whereas physics works mainly from principles and empirical validations, in linguistics ”rationalist arguments” play a central role. By rationalist arguments I refer to those mechanisms

-and it is odd from a theoretical physics viewpoint- that need be postulated at some level of complexity as the starting point of the research. The notion of syntactic structure is one of such arguments, as well as is, in biology, the hypercycle hypothesis for early forms of life. There is also an annoying concept, ”information” which is clearly stated in physics but that it is hard to define in linguistics or biology, due to its entangled connection to concepts as hard as ”meaning” or ”functionality”. I really think that biology is more on the side of linguistics than on the side of physics. Does it mean that it cannot be theoretical? Absolutely not: What we need is to respect the studied object and to provide it with the kind theory it is asking to us. In this way, I vindicate the validity of the contributions of Noam Chomsky, Ferdinand de Saussure, George K. Zipf and John Hopfield to think about general problems of complexity and biology.

A final word is needed. I think that the research program must go back, some day, to physics: both language and life come from the physical world, and, in the end, if we want to really provide a theory of language and a theory of life, we have to connect them to physics. I strongly believe that he need for postulating mechanisms is due to the level of observation we are interested in, and, under this interpretation, rationalist arguments act as the buoy guiding the trajectory of physics. Maybe, with a new form of physics. And, perhaps, to look for this kind of universal connection between sciences is a theological rather than a scientific attitude.

Chapter 1

INTRODUCTION: FOUR KEY QUESTIONS

Language makes infinite use of finite means

Wilhem Von Humboldt

In this chapter the issues addressed in this dissertation are presented. The aim is to put them in the most general and unified way possible, highlighting the motivations and the philosophical problems that lead us to adopt a specific framework and a given scientific attitude. As we shall see, most of our work is mainly based on natural language, but not restricted to it, for it explores properties expected for any complex natural code of communication in its broadest sense.

1.1 Life, Information and language

Living systems are characterized by their complex organization, sustained by a constant flow of matter and energy. Such flows allow them to maintain their internal structure far from thermodynamic

equilibrium. We refer to ”structure” here in a generic way but, generally speaking, it has to do with a spatiotemporal pattern that is not frozen but instead constantly renewed through mechanisms that are themselves the product of evolutionary dynamics. Structures also need to be thought under a multiscale picture: from molecules to populations, life is strongly tied, at each level, with nonequilibrium constraints.

The claim that life follows the laws of physics is a rather obvious one. Living structures cannot escape from the basic laws of thermodynamics. And yet, biological systems depart from physical structures in at least one fundamental property: they actively gather, process and produce information. If we take this in terms of a process of input-output reactions and responses, we could associate to this information flow and its causal implications some form of computation [42, 58] . Information makes a big difference and it does because the relationship between organisms and between them and their environment has been shaped by evolutionary forces. Adaptation in living systems is deeply connected with signals, codes and nonlinear responses, essential to predict the external world. In order to survive, organisms, but also communities and perhaps ecosystems must be able to cope with fluctuations and to do so it must compute.

It was soon realized that information plays a key role and is essential in approaching complex biosystems. The impact of information and computation theories in biology has been big in a broad range of systems, from molecular biology [75] to ecology [2, 26, 79]. It was particularly relevant in the development of the first theories of computation in biology, particularly in relation with brains and their reliability [3, 83, 104].

Perhaps the most sophisticated form of information processing system in biology has to do with one of the major transitions in evolution: the emergence of human language [54, 80]. Human communication represents a major leap in the ways information was exchanged among individuals and specially in terms of its potential generative power. Any other form of communication is, as far as we know,

quite far from our communication system. Complex language represented a turning point for humankind, allowing symbolic thinking to emerge and providing a safe and efficient way of cooperating among individuals within societies. Its impact was huge and we cannot understand our ecological success without incorporating this quantum leap in computational complexity to any complete theory of human evolution.

1.2 Language: A privileged window

In this dissertation we vindicate the key role that human language studies could play within theoretical biology. Moreover, language itself is, to many many researchers, the most difficult problem posed by biological complexity [24]. We thus emphasize that core information-theoretic issues which seem to be crucial for both language and theoretical biology converge in their fundamental set up, but, undoubtedly, they are still open problems.

In this thesis, a structuralist view of biological complexity and its evolution has been taken. By this we mean that we assume that generic mechanisms largely independent upon biological details are at work. These mechanisms (or laws), which do not need to be described in terms of specific microscopic rules, pervade the generation of complex patterns in disparate classes of systems.

The central role given here to human language is not only due to its importance as a key innovation in evolution. Its relevance comes from the observation, at the abstract level, that fundamental scientific problems of human language are also fundamental problems of general theoretical biology. The reason of such general relevance stems from the special location of language at the crossroads between information, cognition, evolutionary innovation and computation. These special status makes any effort made in understanding particular aspects of the problem of immediate relevance to other areas where similar aspects are present.

We can make a tentative list of three key traits of human language that involve such broader problems. Such list will be then addressed below by making five general (open) questions that largely define the path taken in this thesis. The three items are:

1. The presence of generative mechanisms.
2. The question of open-endedness[1].
3. The crucial role of meaning/functionality -which represents a rupture with classical information theory.

A fundamental theoretical hallmark accounting for these phenomena within an hypothetical theory of human language would be a significant step beyond to build a theory[2].

1.2.1 Setting up the problem: Four questions

We explore the three previously outlined fundamental issues through four well defined questions. By formulating specific questions we would like to properly define the boundaries and goals of the work presented here as well as the fundamental results.

The first question has been repeatedly addressed by researchers within both linguistics and evolutionary biology. It is related to the unusual pattern of language acquisition displayed by children. Because of the multiple threads it touches, the acquisition of a mother

[1]Syntactic rules underlying human language are supposed to be able to produce infinite well-formed structures, and thereby the number of potential sentences in a given language is unbounded. This fact is known with several names, like discrete infinity and the mechanisms leading to it are generally summarized with the word recursion. What is clear is that it sets a problem of unbounded increasing in its information content, which can be mapped intuitively to the problem of open-endedness of biological evolution.

[2]Throughout this dissertation we reserve the word theory in its strong version, i.e., as it is known in physics. The debate on whether or not biology -and, moreover, linguistics- can ever accept this kind of approach is open and we do not want to address it here in any definite way.

tongue is actually tied to all the three previous points. Here we present an experimental study of syntax acquisition using network theory. By taking a global, network-level approach, we provide new insights into Question One:

> i) ”What are the empirical patterns we can observe along the evolution of a code displaying an increasing in generative power and complexity?”

This question is the most ”linguistic” one explored here. The results offer a nice picture of the syntax acquisition process and enable to provide tentative answers to the specific problem of language acquisition. Given the central scientific interest of the problem, these observations offer a window to the abstract problem of the emergence of natural, complex codes.

The second question is deeply related to the first point of the list of general problems provided in section ?? is the following:

> ii) ”Can we develop a minimal theory accounting for the abstract process of syntactic structure generation?”

Question Two connects with some of the most well established backbone of theoretical linguistics, but its fundamental presentation goes beyond language. Indeed, a theory accounting for the generation of complex structures based on local operations instead of rewriting rules defines a general hallmark where to study the generative mechanisms. Such framework should be helpful in addressing the emergence of complexity in other information-based structures from prebiotic sets of replicators to collective intelligence.

We follow by studying in depth what is considered the paradigmatic statistical pattern of human languages -and many other complex systems- namely, Zipf’s Law. The third question we have the aim to explore actually involves three sub-questions, namely:

> iii) ”What is the origin of Zipf’s law? Can we provide a mathematical argument accounting for its emergence in

> the framework of information theory? What is the role of Zipf's law when we study complex systems (including human language)?”

Our tentative answers to Question Three are among the major contributions of this dissertation. Given that Zipf's law-like behavior has been reported in many complex systems, from biomass abundance in ecosystems to the distribution of wealth or city sizes -see [84] and references therein-, its scope goes far beyond human language. The crucial point here is that an explanation in the framework of information theory as the one presented here is free of specific mechanisms and, therefore, of general applicability. We first develop an abstract mathematical apparatus and then demonstrate that, when applied to the study of the emergence of such a law in human language, Zipf's law is found to be the expected one.

Question Four involves one of the most elusive aspects associated to generative mechanisms and statistical patterns from an information-theoretic viewpoint. It deals with the intrinsic lack of any form of meaning, functionality or referentiality. The abstract picture that can be extracted from these meaning-free approximations is necessarily incomplete. We contribute to fill this gap by exploring the problem of the inclusion of a simple referential system in a standard measure of information. We restrict the problem to the framework used in current models of autonomous communicating agents -see [69] and references therein-, where the emergence of communication is studied. Within this framework, we ask:

> iv) Can we quantify, in terms of information theory, the degree of conservation of the referential value in a given communicative exchange?

This quantification leads us to explore the interesting consequences of the conservation of referentiality in classical communication schemes. As will be shown, in this dissertation we solve an apparent paradox reported in previous models of the emergence of communication.

These four questions define the starting point of the contribution provided here. The order by which we presented them is not arbitrary: They define an increasing level of generalization. Indeed, the way by which the first question is addressed can be considered, in many -but not all- levels, to be linguistic specific. The second one is more general but still strongly tied to linguistics, for it is clear that all the abstract machinery comes from language studies. The emergence of Zipf's law is a more general question, since it includes human language but also many other complex systems. The fourth problem is still more general, for it addresses an essential of classical information theory and proposes a way to fill it.

1.3 Generative mechanisms and patterns of unboundedness

Several comments are in force, in order to clearly accommodate this dissertation to a broader research program. Specially controversial is the scientific status of generative mechanisms. The study of generative mechanisms and the question whether or not such mechanisms are able to define a system with an unbounded information capacity quickly collides with the problem of empirical validation. Indeed, real data is necessarily finite and the potential rules underlying experimentally established regularities -if any- are intrinsically unaccessible. Therefore, we have to be able to identify the footprints of unboundedness through statistical analysis over real data.

The presence of generative mechanisms must be rationally postulated as the simplest solution to explain empirical patterns, and by demonstrating that without them, paradoxical situations occur. To this end, we have to refer to the paper [81] where Chomsky and Miller suggested that (finite) statistical analysis will always lead to incomplete descriptions of human syntax. The reason was the identification of several phenomena that lead to postulate structure-dependent syntactic relations regardless the linear distance between sentence's

elements [20, 21]. Without a rational ansatz concerning the presence of a mechanism able to generate structure, statistics is necessarily unable to identify the generative mechanisms. The core of syntactic theory, is argued, must avoid the statistical analysis and it is necessarily tied to a computation/theoretic viewpoint where the rules are validated through a controversial empirical proof based on the grammaticality criteria of speakers having a given language as the mother tongue -see [19, 20] defending such an empirical criteria and [55] to understand the objections against it. I consider reference [81] paper as one of the most intriguing paper of general science of the XXth century, for it clearly sets a nowadays still unsolved problem for the all-powerful statistical inference as the method to validate data and construct theories.

The hallmark defined by Miller and Chomsky is assumed in this dissertation, thereby going from statistics to set and computation theory and the other way around. Specifically, the information-theoretic study of the statistical patterns displayed by a system having internal generative rules is complemented with a set-theoretic study of the minimal properties expected for a generative mechanism in the natural world, along the footsteps of the so-called ”Minimalist Program for a Linguistic Theory” [22, 23]. These two approaches are not mutually exclusive: Instead, we need to clearly state the scope of each approach. In the following section we discuss the approaches we choose to afford questions i)-iv), which include both statistics and computation (set) theory. The main divergence with Miller and Chomsky's philosophical program is that we do believe that an integration of the two levels of study is possible, although accepting the limitations of each approach and the different levels of study they represent.

1.3.1 Language acquisition

According to question i), can we observe the emergence of a complex communicative system in order to take empirical data from it?

This issue directly concerns the study of actual statistical patterns of complex communication systems. The exploration of the language acquisition process is one of the most fascinating problems of contemporary science, and it is a paradigmatic example of the emergence of combinatorial complexity in the real world.

It has been known for a long time that the process of language development in infants is a highly nonlinear one. In a nutshell, after an apparently smooth process of increased vocabulary growth where single words are at some point replaced by pairs of words, a sudden change occurs around the second year that shifts the protolinguistic pattern -as named by Bickerton, [11]- into a complex, grammar-driven organization of language [86, 91]. In other words, we jump from a two-word phase into a phase where the rules of language generation seem to be fully at work. This transition is far from trivial and has received considerable attention. It has also been very controversial. Once again Chomsky was at the origins of the controversy when he claimed that, as it happens with other organs of the body, which have no need of "being trained" to perform a given function, there should be something innate hardwired in human's brains ready to develop language [20]. The language acquisition device would "explain" the apparently innate facility of acquiring a language in early life.

The problem of language acquisition in children is unfortunately also plagued with qualitative and sometimes ideological arguments. A considerable debate has been developed over the years in terms of the validity or absurdity of his language device. Inventories of words, well-defined statistical explorations and other quantitative treatments have been used in order to characterize the transition. It is interesting to mention that available studies concerning brain development in children and its correlations with speech reveal a range of activity patterns changing over time (figure 1.1a) but no dramatic transitions. Unfortunately, none of these studies could possibly end the debate. Such approximations have been important and helpful in defining some basic background, but they have largely been ig-

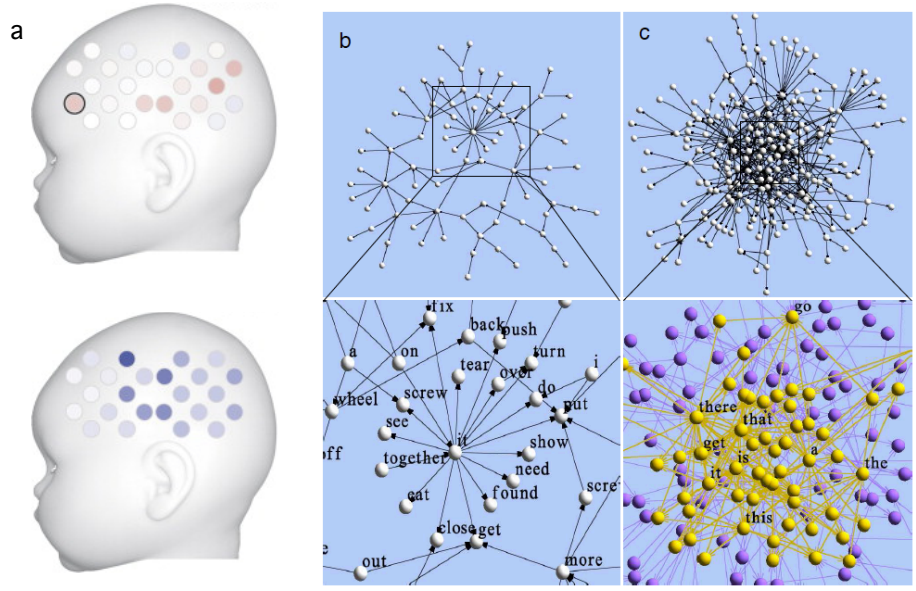


Figure 1.1: The brain experiences great changes within the early life of humans. Children brains have been studied (a) in searching for changing patterns of activity under given tasks. Here differences in speech processing cortical maps are depicted for the left hemisphere of three-month (top) and ten-month old infants. Here the color scale is related to the response activity in controlled responses to speech stimuli -from [56]. Differences have been reported across all childhood from different sources, although none seems to involve a dramatic change in network patterns correlated to syntax network architecture. Network patterns obtained from the aggregation of syntactic relations of child's utterances during different periods of language acquisition. The first network (b) belong to the so-called 2-words stage, and it is worth to note the tree-like structure of the net (except for a few crossings). The second network (c) is obtained after the so-called syntactic spurt, and its topology is both quantitatively and qualitatively different. Both in (b) and (c) we display the connected component as well as a zoomed subnetwork in order to display some of the specific items contained at the core of each graph.

noring something essential to the nature of language structure: the patterns of word-word interactions. In particular, the network organization that can be extracted by systematically looking at all the relationships among words in a given corpus or conversation provides a much better and unbiased characterization of language structure than simple inventories [98].

We attacked the problem by looking carefully at syntactic networks reconstructed [27] from available data sets on CHILDES Database [14], [15] -available at http://http://childes.psy.cmu.edu/-, using strong indicators of network complexity, such as small world measures, motif patterns, degree distributions or the chromatic number. With this novel type of analysis, it was possible to reveal a previously unreported quantitative transition from non-syntactic to syntactic communication that seems to give support to the presence of some predefined ”language device” triggered at a critical point of the acquisition process [36].

As we shall see, the observed patterns are consistent with the emergence of a qualitatively different system of syntactic structure generation. However, as we pointed out above, the finite nature of statistical data is necessarily incomplete to conclude that the generative mechanisms are able to produce syntactic structures having an unbounded size. The observed patterns are consistent with such an hypothesis, but the rationalist starting point -as the one provided in [81] is needed for this conclusion. We do not close at all this debate, the only aim we have here is to present the patterns one can expect if the hypothesis of the emergence of an unbounded generative mechanism triggered at some point from some innate endowment holds.

As far as we know, this is the first quantitative analysis of syntactic change using a global, network-level perspective os syntax. In this context, and in spite of its shortcomings, it is a clear evidence for the presence of a nonlinear, sharp and qualitatively nontrivial phase change. Such transition cannot be explained from any typical network growth mechanism involving percolation-like phenomena.

1.3.2 The minimal generative system

The tempo and mode of language acquisition dynamics opens a number of important questions. Given that the transition that occurs around the two-year boundary involves a dramatic jump from two words-structure to unboundedness (always taking these claims cautiously), it indicates that the generalization made by the brain out from a limited repertoire of examples is likely to be obtained in terms of some basic generative system able to cope and generate recursive structures [54, 80, 92, 93]. Recursion allows us to build an infinite number of posible sentences, by properly nesting subsets of words within others under syntactic constrains. The study of recursion is a rather difficult problem, since it goes beyond combinatorics (something often used as equivalent in some language models, when it is not) and touches the deepest grounds of language: the generative rules leading to complex language and their cognitive origins. Although no specific response is given on to how these generative rules are implemented in our brains, this dissertation presents a novel formal approach to the generation of syntactic structures [28, 45] and that provides, we believe, a much better framework where some of our key questions might get an answer.

Following the basic scheme of the so-called Minimalist program for a linguistic theory we study question ii) by exploring a set-theoretical approach to the generation of syntactic structures based on the so-called ”merge” operation. Roughly speaking, this basic operation allows the combination of two syntactic objects into a new syntactic unit. It allows recursion, since the same operation can be applied to the resulting output and so on. In other words, objects combined by this operation are either lexical items or units already created by merge.

Merge operation has been identified as the key formal innovation that could explain the emergence of modern human syntax [12, 54]. By identifying this operation with the well known set-union operation, we propose a minimal theoretical backbone where a minimal

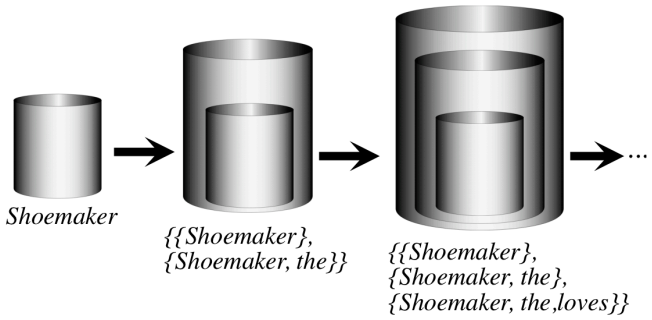


Figure 1.2: Syntactic structures are nests, not chains of elements. In this picture we show how a nesting algorithm works using the metaphor of the physical recipient: In the example proposed in the main text, the first element to be merged is a small recipient embedded in a bigger one -which corresponds to the syntactic constituent $\langle \text{the}, \text{shoemaker} \rangle$. This operation can be hypothetically expanded in an unbounded way generating arbitrarily large nested structures.

generative theory based on merge is accommodated. In this approach, ”recursion” is formally stated as ”nestedness”, and the fundamental entities of syntactic theory are nests, i.e., families of subsets of a given set ordered by inclusion [66, 101].

To grasp the intuitive idea, let us briefly revise, in a simplified way, how it works for the following sentence, at the level of constituent structure:

”Mary loves the shoemaker”

under our theoretical framework, the syntactic derivation of this sentence starts by defining the set

$$A = \{\{\text{Mary}\}, \{\text{loves}\}, \{\text{the}\}, \{\text{shoemaker}\}\}$$

The derivation is obtained through the following algorithm:

$$\begin{aligned} M_0 &= \{\text{shoemaker}\} \\ M_1 &= M_0 \bigcup \{\text{the}\} \\ M_2 &= M_1 \bigcup \{\text{loves}\} \\ M_3 &= M_2 \bigcup \{\text{Mary}\} \end{aligned}$$

The syntactic object structuring such a linguistic production, N is:

$$\begin{aligned} N &= \{M_0, M_1, M_2, M_3\} \\ &= \langle M_3, \langle M_2, \langle M_1, M_0 \rangle \rangle \rangle \\ &= \langle \text{Mary}, \langle \text{loves}, \langle \text{the}, \text{shoemaker} \rangle \rangle \rangle, \end{aligned}$$

where the second equality comes from Kuratowski's definition of ordered pair, namely $\langle a, b \rangle = \{\{b\}, \{a, b\}\}$. We observe that N is a nest over A, since

$$M_0 \subset M_1 \subset M_2 \subset M_3.$$

This provides a formal starting point from which a mathematically rigorous theory of generation of syntactic structures can be built. The conceptual innovation concerns precisely the mathematical characterization of derivational framework adopted, instead of the classical one provided by rewriting rules, commonly used in the theory of formal grammars. This conceptual shift reinforces the role of how structures are generated, since merge operation is hypothesized to have biological support.

Although the "existence" of such innovation is the object of a hot debate, and our position is far for being dogmatic: The scientific attitude here has been to provide a solid mathematical ground to what seems to be the minimal generative system supposedly able to represent the complexity of human syntax -and in general, a plausible natural mechanism to generate unbounded codes. As we said above, how this is physically implemented is a fascinating problem, but lies far away from the scope of this dissertation.

1.3.3 Zipf's law

Zipf's law is a prominent statistical law that seems widespread in all languages. It takes the name of the linguist George K. Zipf, and it states that given some corpus of natural language, the frequency of any word is inversely proportional to its rank [108]. Specifically, if we rank all the occurrences of words in a text from the most common word to the less common one, Zipf's law states that the probability $p(s_i)$ that in a random trial we find the i-th most common word $(i = 1, ..., n)$ falls off as

$$p(s_i) = \frac{1}{Z} i^{-\gamma}, \tag{1.1}$$

with the exponent, $\gamma \approx 1$, and being Z the normalization constant, i.e.,

$$Z = \left(\sum_{i \leq n} i^{-\gamma} \right). \tag{1.2}$$

In other words, the most frequent word will appear twice as often as the second most frequent word, three times as often as the third, and so on.

This law is also well known in many other fields and has received great attention within both statistical physics and theoretical biology due to its commonality [6, 13, 46, 71, 76, 88, 96, 108]. It has been observed in a plethora of systems of different nature and scale, including biomass distributions, city size profiles or extinction events, to cite just a few see [35] and [84] and references therein. This law tells us that the vast majority of words in any corpus (and in language use) are rare, whereas a few words will be extremely common. The pattern of appearance is rather robust and seems to tell us something of great relevance. And yet, in spite of this nontrivial pattern, quantitative theories of language evolution have been largely ignored within linguistic theories.

From the mathematical viewpoint, Zipf's law appears to be a rather special distribution, and, in spite of its simplicity, it has a

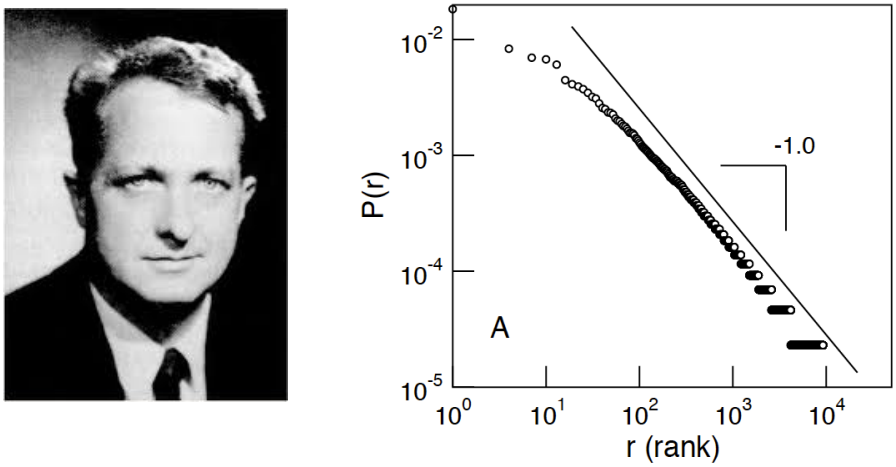


Figure 1.3: George Zipf (right) first reported about the widespread presence of a special scaling law with a well-defined scaling exponent. This law can be observed in human language (b) and many other contexts. The plot displayed at left shows the frequency distribution of words from a written text in log-log scale. The linear trend in this diagram indicate the presence of a power law.

great amount of ”undesirable” properties. Its huge heterogeneity and scaling form suggest that some non-trivial phenomenon pervades it. Theoretical approaches to Zipf’s law have tipically considered specific microscopic mechanisms, and many different types have been found -among others, see [7, 41, 43, 52, 60, 74, 77, 78, 99, 85, 103, 107]. Since they differ considerably in relation to their specifications, the universal character of the law strongly points towards some sort of fundamental cause beyond specific, given mechanisms. Guided by this conjecture, it was possible to develop a novel framework, based on a few key concepts from information theory, showing that indeed a rule-free theoretical derivation of Zipf’s law can be constructed [31, 35].

The study of question iii) begins with the definition of the mathe-

matical apparatus we built to derive the general conditions that lead a given system to exhibit Zipf's law. This mathematical hallmark is based on the explicit formalization of the basic properties of a system growing between order and disorder. Interestingly, this formalization revealed compatible with the formalization of the so-called least effort hypothesis, conjectured by G. K. Zipf as the responsible of the emergence of such a distribution in human language [108]. Zipf's arguments emphasize the role of cooperation as the crucial feature organizing the whole code. Let us be more concise: Zipf's picture involves two agents, to be identified as the coder and the decoder. The coder sends information from an arbitrary environment to the decoder agent, and in turn, this latter agent inferrs the behavior of the environment from the information provided by the coder one. Following Zipf's arguments, two opposite, non-cooperative strategies can be defined:

1. The first strategy implies the least effort for the coder: whatever the behavior of the environment, only one signal is sent. This strategy implies that all the effort in interpreting the surroundings is led to the decoder agent.

2. The second one is defined by assuming that the coder agent codifies every event of the environment with a specific signal, thus providing to the decoder one an unambiguous code, thereby implying that this latter agent needs no effort to infer the environment behavior.

This scenario describes a tension between the coder and the decoder, a tension that can be solved by a cooperative regime in which the amount of additional information needed by the decoder to completely reconstruct the environment is equal to the one provided by the coder. Although Zipf's arguments were mainly intuitive, its mathematical abstraction is quite straightforward, standing out the effort made in [52] to determine a rigorous mathematical framework where Zipf's intuitions were well established. A less formal approach, based on numerical simulations was provided in [43] and in [102].

Previous works addressing this topic were concerned with the mathematical formalization of the cooperative regime between communicating agents as the sole source of the emergence of Zipf's law. As we shall see, it is no difficult to demonstrate that this constraint alone does not account for the emergence of Zipf's law, but actually can lead to an arbitrary family of probability distributions, if we accurately tune their parameters. To obtain Zipf's law as the unique solution to the equations obtained to describe the cooperative regime we need another ingredient, evolution, which is an unavoidable feature of the systems exhibiting Zipf's law. Indeed, we assume that the code is not static but that it grows through an arbitrary time scale, and that the growing process is governed by the so-called Minimum discrimination of information Principle (henceforth $MDIP$) [72], which implies that the configuration of the code through different stages of the evolution is minimal, in agreement to the least effort intuitive idea. It is worth to note that Zipf's arguments completely ignore this fundamental ingredient. The $MDIP$ has an interesting mathematical consequence: It quantifies the path dependence inherent in any evolutionary process, understood in its broad sense. The bill we have to pay is that we no longer work with an stochastic object -the code- whose dimension is stable, but instead, its dimension grows unboundedly during the evolutionary process. We emphasize that the fundamental character of such a variational principle could have a wide role in non-equilibrium statistical mechanics[3].

The rigorous demonstration, for the first time, of Zipf's conjecture -see [31, 35]- as the origin of the law having his name in a framework fully embedded in information theory is an achievement by itself. But, concerning one of the main common threads of this dissertation, why this statistical pattern has a potential role in the identification of unbounded information systems? As we shall see, Zipf's law has a very special role in the zoo of probability distributions: Zipf's law is the distribution that defines the border between bounded information codes and the unbounded ones.

[3]See the discussion provided in chapter 2, section 2.1

1.4 The Problem of Referentality

The next point of the list proposed above considers the uncontroversial appreciation that life -and language- manages information in a meaningful way [58, 106]. This has been made more concrete through question iv). As we pointed out in the list, this implies a rupture with classical information theory. The reason is clear: Information theory is formulated by explicitly ruling out the meaningful content of messages [49, 50, 51]. This is odd if we want to directly use the framework of information theory to build a theory of biology or language. This lack of the semantic element was not a problem from the engineering viewpoint, but implies a clear limitation on the extent of the conclusions and applicability of the formal technology provided by such a theory, which is commonly neglected. Early approaches to overcome this problem tried to introduce the logical/semantic component in the general framework of the theory [8].

A fundamental conceptual shift in the area arrived in the sixties, with the definition of the absolute information content [17, 68, 100], which left behind the relative nature of the Shannon information, and established a more powerful conceptual background for information theory. This idea of absolute information content (also known as "Kolmogorov complexity") proved extremely useful to deal with a wide variety of theoretical problems, from the philosophy of mathematics to the physics of computation. Additionally, it enabled researchers to work with a more clearly defined concept of complexity. And, quite crucially, it also defined the starting point for a theoretical unification of computation, complexity and information theories [82]. Nevertheless, it was soon pointed out that several issues, intimately linked to biological systems or linguistic structures (e.g.: the emergence of increasingly complex -but non-random- molecular structures, the functional role that these may play in the organization of metabolic processes, in their stability and robustness, in their longer-term evolution,...) did not find so easily an answer or a possible treatment within this new foundational theory and, therefore,

further work had to be carried out in that direction. A specially interesting attempt to account for the emergence of complexity in terms of information theory, through the notion of effective complexity was presented in [47] (a young, promising concept whose impact is still difficult to evaluate).

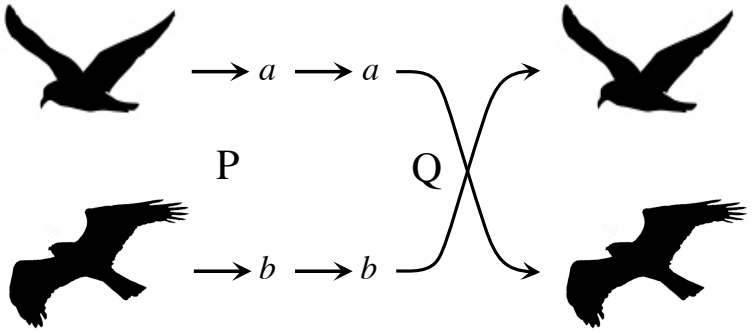


Figure 1.4: An example of the potential consequences of an absence of referential value. Two opposed events (detecting an individual of the same species or a predator) are coded in some way by **P** and deterministically decoded by **Q**, shifting the referential values of the signals. In terms of classical information theory, this system could display maximum mutual information. However, if functionality of the communicative exchange is supposed, the selective consequences of this scenario -where friends and enemies can be mistaken- is completely disastrous. A "perfectly wrong" identification of a predator as a friendly neighbor is a deadly one.

Beyond this complexity-theoretic concepts, one has to be aware that we are facing the general problem of meaning. In the case of human language, Wittgenstein's works on the fundamental problem of meaning, to whom "there is no room for a systematic theory of meaning" [70, 105] represent one of the most clear examples of thinking at the edge of scientific inquiry[4]. Even the fundamental question re-

[4]It would be interesting to know how these ideas and metaphors can be trans-

mains unsolved and maybe, has no general solution. In this context, several studies explored the problem by assuming a simplified conception of the referential value of a given signal, as shown in the toy example provided in figure 1.4 [61, 87, 69]. These studies are based on different formalizations of the dual nature of the communicative sign, a fundamental structure of any communicative signal postulated by the swiss linguist Ferdinand de Saussure [94]. Saussurean duality states that any communicative sign is composed by the signifié and the signifiant, which can be roughly interpreted by considering the communicative sign as a pair -instead of a single element- containing a given signal and its associated referential value, which is taken among the members of a set of potential referents. It is worth to emphasize that it is an extreme simplification of the Saussurean concept of duality of sign, but enabled researchers to work beyond the structure of signals and codes provided by information theory. Specifically, this conception of the communicative unit lead researchers to define game-theoretic based models of the evolution of communication inside the population of autonomous, abstract agents which are able to communicate among them [61, 87, 69]. The contribution of Ferdinand de Saussure crucially identified one of the core differences of natural communication systems and the artificial ones -where the semantics does not emerge, but it is explicitly defined.

Here the problem of referential value is explored by taking, as the formal starting point, the probabilistic framework proposed by several authors [61, 87] see also [69]. The first issue addressed here is the emergence of non-self consistent agents in an evolutionary context. Given that an agent has a dual nature -i.e., it has a coder module and a decoder module- we can easily conceive an agent which is not self-consistent. By non self-consistency we refer to those agents that, although able to successfully communicate with other agents, their internal coder/decoder configuration are not compatible and, thus, the agent does not understand itself. This has been stated as a para-

lated into some fields of biology, particularly to the well-known problem of the mapping between form and functions (or genotype-phenotype).

dox that needs to be specifically fixed, through the incorporation of additional assumptions in the definition of agents [61, 69]. In [34] we show that evolution solves this apparent paradox, or more, specifically, makes its emergence unlikely.

The second issue is more general. Accepting that the above observations put into question the current form of Shannon information to deal with this kind of problems, we try to expand the notion of mutual information to account for these kind of undesirable situations. To account for the amount of ”well referentiated bits”, we expand the classical Shannon's mutual information with a referential parameter, which is a ratio of well ”referentiated” bits against total bits needed to describe the system. This information-theoretic measure has interesting consequences and grasps -even in a very simple way- the role of some kind of referential value for the communication schema.

Chapter 2

THEORETICAL FRAMEWORK

I don’t care about the content of the message.

Claude E. Shannon

Information theory was the brainchild of Claude Shannon, who defined the foundations of this field in a groundbreaking work that deeply influenced science, philosophy and engineering. It also provided a novel and powerful quantitative framework for connecting evolution, information and physics. Its impact on science and society were huge. It has permeated our understanding of how complexity is defined, how it evolves and emerges our of equilibrium. Although the mathematical background of this dissertation includes set theory, graph theory and information theory, it is clear that the latter acts as the common thread underlying all the results provided here.

Other key players were Andréi Kolmogorov and Gregory Chaitin, whose work has been also fundamental in understanding some key aspects of computational complexity and its limits. In this context, although the work presented here is grounded in several frameworks, from set theory to graph theory, information theory will be our main formal resource. This chapter presents, in a nutshell, the basic concepts of information theory which underlie an important part of the

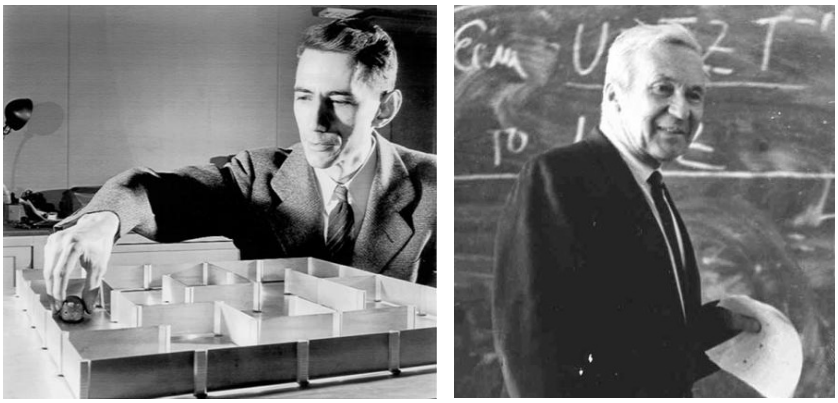


Figure 2.1: Claude Shannon (left) the founding father of information theory, provided the mathematical basis for a general framework of information coding and decoding through channels. The theory provides a powerful approach to the problem in both natural and man-made systems. Other researchers, such as the russian genius Andréi Kolmogorov (right) also contributed to this area, allowing to connect complexity and computational complexity in mathematical terms.

dissertation. It is not intended to be an exhaustive introduction to this branch of mathematics, and its technical backbone -e.g. the axiomatic derivation of entropy, or the Asymptotic Equipartition Property- is deliberately ommited, reinforcing the intuitive understanding of the presented concepts and highlighting their possible relations and limitations and including personal considerations.

For a deep understanding of the section related to basic definitions, I refer the interested reader to the original papers of C. E. Shannon [95] or the standard textbooks [5, 37, 63]. Its location within the mathematical theory of probability is well developed in [67] also in [5]. The Kullback-Leibler divergence is presented in detail in [72] and in [37][1].

[1]It is noteworthy that both textbooks locate such a concept at the core of

A brief presentation of computational complexity is provided in the framework of Algorithmic Information Theory. For the sake of fluency and simplicity, Turing Machines are not explicitly introduced. We will take as a definition that a Turing machine is an abstract machine able to perform any computation that can be described through an algorithm -a finite set of logically consistent rules that can be executed without any creativity or intelligent contribution[2].

In relation to Algorithmic Information Theory, we refer the interested reader to the seminal works [17, 18, 68, 100]. The standard textbook on this topic is [82], but additional introductions can be found again in [37] and [49, 50] where the connection between Kolmogorov Complexity and Entropy is discussed. Criticisms to the ontological value of Kolmogorov Complexity as the actual measure of complexity are given in [47].

2.1 Entropy and information measures

Throughout this section we will consider the following scenario: Let us suppose we have two random variables X, Y taking values in sets $\mathcal{X}, \mathcal{Y}$, with associated probability functions $p(x), q(y)$, conditional probabilities $\mathbb{P}(x|y)$, $\mathbb{P}(y|x)$ and joint probabilities $\mathbb{P}(x, y)$. We will suppose that $(\forall x_i \in \mathcal{X})(p(x_i) > 0)$ and the same applies to $Y, \mathcal{Y}$ and q.

information theory. In opposition, Shannon's ealry works [95], as well as [5, 63] pay a few or no attention to this concept.

[2]I personally recommend the interested reader [10, 38, 57] for their rigorous presentation of Computation Theoretic issues and how the set up the central role deserved to Turing machines in the foundations of Mathematics and Physics.

2.1.1 Shannon Entropy

The Shannon entropy or uncertainty associated to the random variable X, $H(X)$, is defined as:

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x). \tag{2.1}$$

(Throughout this dissertation $\log \equiv \log_2$, unless the contrary is indicated). We observe that eq. (2.1) is actually an average of $\log(1/p(X))$ among all events of $\mathcal{X}$, namely:

$$H(X) = \mathbb{E}\left(\log\frac{1}{p(X)}\right),$$

were $\mathbb{E}(...)$ is the expectation or average of the random quantity between parentheses[3]. The extreme $H(X) = \log|\mathcal{X}|$ is achieved when

$$(\forall x \in \mathcal{X}),\ \ p^*(x) = \frac{1}{|\mathcal{X}|}. \tag{2.2}$$

Some remarks are needed in order to understand the scope of equation (2.1). Shannon's early interpretation of $H(X)$ was related to the amount of informative richness we can extract from a given random variable. Or, in other words, if the random variable X acts as a potential information source, $H(X)$ is the amount of bits it can provide. Consistent to this interpretation, let us suppose that we encode the outcomes of X with a binary code. The Shannon-Fano-Elias theorem for optimal coding [37] states that the minimal length, in bits, of $x \in \mathcal{X}$ is:

$$l^*(x) = -\log p(x) + \mathcal{O}(1). \tag{2.3}$$

[3]As a concave function, the entropy satisfies the so-called Jensen's inequality, [37] which, in this specific case, reads:

$$0 \leq \mathbb{E}\left(\log\frac{1}{p(X)}\right) \leq \log\left(\mathbb{E}\frac{1}{p(X)}\right) \leq \log|\mathcal{X}|,$$

Thereby, the average minimal length, in bits, of the binary code generated from the coding of X is:

$$\sum_{x\in\mathcal{X}} p(x)l^*(x) = H(X) + \mathcal{O}(1). \tag{2.4}$$

Another, most fundamental view of $H(X)$ was provided by Khinchin, and stems from its mathematical interpretation within the core of the theory of stochastic phenomena. Specifically, $H(X)$ (up to a multiplicative constant) is the only function satisfying the so-called Uncertainty Axioms, which legitimates $H(X)$ as the only function able to quantitatively evaluate the intrinsic uncertainty of a random variable. We refer the interested reader to [67] and [5] for the details of this interesting interpretation.

Let us end this summary of $H(X)$ by highlighting the formal equivalence of $H(X)$ to Boltzmann entropy S, which makes $H(X)$ a suitable measure of entropy from which Statistical mechanics can be constructed. This connection was rigorously established in the Jaynes' influential works [62]. Roughly speaking, Jaynes derived the results of equilibrium statistical mechanics from the maximization of the entropy subject to the constraints known by the observer[4]. Therefore, if we know something about the system -e.g, the average energy of the molecules in a gas- the most expected configuration of the system is the one that maximizes the remaining ignorance -thus, the degree of uncertainty- over the system. Quoting Murray Gell-mann, Jaynes' principle can be summarized as follows: "don't say you know more than you know " [47].

Jaynes' work introduced a simple variational principle where entropy maximization was achieved under known constraints. Variational principles have a central role in physics [4, 48]. For example, the time evolution of a mechanical system is governed by the minimization of its Lagrangian functional [4]. However, the scope of the

[4]These constraints are generally known as moment constrains and include averages on observables, like the internal energy -see [62, 64, 89].

principle of entropy maximization revealed non-satisfactory to explain the origin of power-law configurations and, in general, systems' configurations out of equilibrium. Indeed, non-equilibrium systems produce entropy, and they are below its maximum entropy configuration.

To work with evolving systems out of equilibrium we need another variational principle. In this dissertation, a particular non-equilibrium problem is successfully solved by using the Kullback-Leibler divergence as the functional to be minimized along system's evolution. This functional is presented in the following section.

2.1.2 Kullback-Leibler Divergence

The "Kullback-Leibler (KL) Divergence", "Relative Entropy" or "Information gain" of the distribution q with respect to the distribution p, $D(p||q)$, is defined as:

$$D(p||q) \equiv \sum_{x \in \mathcal{X} \cap \mathcal{Y}} p(x) \log \frac{p(x)}{q(x)}. \tag{2.5}$$

$D(p||q)$ can be understood as a measure of distance (although it is not symmetric) between two distributions. Here $q(x)$ is a reference probability distribution. Its meaning (see below) is that, on maximizing entropy, $p(x)$ is equal to $q(x)$ in the absence of any constraints.

The KL divergence evaluates, in bits, the amount of extra information needed to perfectly describe a random object whose real associated probability distribution is p by proposing a code with associated distribution q. In this way, if we use q to sample p, we will need at least

$$H(X) + D(p||q) = -\sum_{x \in \mathcal{X}} p(x) \log q(x)$$

bits to perfectly reconstruct X. The above quantity is often referred to as the cross entropy of q with respect to p. A note of caution

is needed. Although $D(p||q)$ is the minimal information needed to reconstruct p from q, it tells us nothing about the specific realizations of X and how they are related to Y. In this way, one could have $p = q = p^*$, thereby having $D(p||q) = 0$ but, at the same time, have no correlation between X and Y. In this case, we can say that we can sample perfectly p using q, but we cannot say anything about a given value of of X knowing a given value of Y or viceversa[5].

The relevance of $D(p||q)$ within the framework of information theory is second to none. Moreover, the existing relation between $H(X)$ and $D(p||q)$ clearly states that, at least, it has the same status of entropy. Let p^* the probability distribution defined in eq. (2.2). Then:

$$H(X) = \log |\mathcal{X}| - D(p||p^*). \tag{2.6}$$

In the general case, it can be shown that [47]

$$|H(X) - H(Y)| \geq D(p||q). \tag{2.7}$$

But we can go further and say that $D(p||q)$ is, in fact, a generalization of $H(X)$. Indeed, eq. (2.6) defines the entropy as a measure in terms of the distance between our distribution and the absolute disorder. This is actually a particular case of the Kullback-Leibler divergence, since the reference distribution in this functional is left open, as depicted in eq. (2.7). This is why it is often referred as relative entropy [37].

This relative nature opens the possibility -and this is a conjecture- to use this functional as the central piece of non-equilibrium phenomena, and to interpret its minimization along the different stages of the evolution of an statistical ensemble as a kind of minimum entropy production principle. In this way, it has been shown that the

[5]This represents the first important divergence between the information-theoretic framework proposed in this dissertation to study the emergence of Zipf's law [31] and the proposal made by [52] and [53]. In our approach, we take into account this issue, whereas in the works of Harremoes and Topsøe they only consider divergences between distributions and, therefore, the coding processes is deliberately omitted.

minimization of $D(p||q)$ in systems whose evolution is describable through the Fokker-Planck equation [90] leads to time-dependent solutions. Furthermore, as we pointed out above, in this dissertation we show that the unicity of Zipf's law as a solution of a general class of problems of of entropy restriction [52] can be justified assuming that the transition between successive stages of the evolution is governed by the minimization of their relative entropy, in the same way by which two successive stages of a non-equilibrium system are governed by the principle of minimum entropy production [73].

2.1.3 Conditional and Joint Entropies

The "conditional entropy" or "conditional uncertainty" associated to the random variable X with respect to the random variable Y, $H(X|Y)$, is defined as:

$$H(X|Y) = -\sum_{y\in\mathcal{Y}} p(y) \sum_{x\in\mathcal{X}} \mathbb{P}(x|y) \log \mathbb{P}(x|y). \tag{2.8}$$

$H(X|Y)$ is the uncertainty remaining on the system composed by the two random variables X, Y when we have access to the knowledge provided by Y. As expected,

$$H(X|Y) \leq H(X) \tag{2.9}$$

The uncertainty in recovering the specific values of X from the knowledge of Y received the name of noise in classical information theory.

We can also define the joint entropy among two random variables X, Y, written as $H(X, Y)$:

$$H(X,Y) = -\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} \mathbb{P}(x,y) \log \mathbb{P}(x,y). \tag{2.10}$$

The joint entropy is the minimal amount of bits needed to properly describe the composite system X, Y. It is symmetrical, i.e.:

$$H(X,Y) = H(Y,X). \tag{2.11}$$

The correct interpretation of both $H(X|Y)$ and $H(X,Y)$ is crucial to understand the scope and significance of the mutual information, to be defined in the next subsection. $H(X|Y)$ and $H(X,Y)$ are connected by the following equality:

$$H(X,Y) = H(Y) + H(X|Y). \tag{2.12}$$

Furthermore, the upper bound on $H(X,Y)$ is found by application of eq. (2.9), namely:

$$\max\{H(X), H(Y)\} \leq H(X,Y) \leq H(X) + H(Y) \tag{2.13}$$

The case $H(X,Y) = H(X) + H(Y)$ refers to a composite system where no causal relation can be inferred from the behavior of X over the behavior of Y. In this case -the case of complete uncorrelated system-, $H(X,Y) = H(X) + H(Y)$ is the explicit realization of the so-called Additivity Axiom for the measure of uncertainty [5, 67].

2.1.4 Mutual Information

How can we quantify the relations existing among the realizations of X against Y, or viceversa? Does the behavior of X determine in some way the behavior of Y? If so, the system will have some degree of predictability. If not, X and Y will be completely independent random variables. This kind of questions have quantitative answers thanks to the so-called mutual information.

The "mutual information" $I(X:Y)$ among the two random variables X, Y can be defined in several, equivalent ways [5, 37, 95]. The first we consider is the following one:

$$\begin{aligned} I(X:Y) &= D(\mathbb{P}(x,y)||p(x)q(y)) \\ &= \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} \mathbb{P}(x,y) \log \frac{\mathbb{P}(x,y)}{p(x)q(y)}. \end{aligned} \tag{2.14}$$

This first equality defines the mutual information from the Kullback-Leibler divergence, and has to be interpreted in the following way:

The mutual information grasps the distance in bits obtained from the observation of pairs $\langle x, y\rangle$ of our composite system X, Y against the expected distribution one could expect by pure random associations.

If the divergence is strong, we infer that the system has important correlations among the two involved random variables which results in a high degree of predictability of the behavior of Y knowing the behavior of X. Notice that, according to the interpretation of the Kullback-Leibler divergence provided in the above section, one can infer that $I(X:Y)$ is the amount of deviation of the behavior of the composite system X, Y against the null hypothesis. This is why the Kullback-Leibler divergence is often referred as information gain.

The second definition is the standard one and the most intuitive of all alternative definitions:

$$I(X:Y) = H(X) - H(X|Y) \tag{2.15}$$

This definition clearly separates the behavior of the source and the noise. In this definition, $H(X)$ is the total amount of informative richness available from the source. The noise term $H(X|Y)$ evaluates the degree of uncertainty introduced by the process of reversion: If, whatever the value of Y, we can unambiguously determine the value of the source, then $H(X|Y) = 0$ and the mutual information will be maximum. On the contrary, if $H(X|Y) = H(X)$ the process is completely noisy, since all information is destroyed. In the latter scenario, the behavior of Y can no longer be predicted from the knowledge of X.

Using identity (2.12) we reach another, interesting formulation of the mutual information:

$$I(X:Y) = H(X) + H(Y) - H(X,Y). \tag{2.16}$$

In this definition, mutual information is computed by taking the information richness of the two random variables independently against the term accounting by the global information contained needed to describe the whole system, namely $H(X,Y)$. In this way, it is intuitive that if the behavior of Y can be predicted from the behavior of

X, then the overall of bits needed to describe the composite system X, Y will be reduced, thereby increasing $I(X : Y)$, since $H(X)$ and $H(Y)$ are independent measures.

The close examination of this measure raises several questions about its range of applicability. Needless to say, its formulation is one of the most important intellectual achievements of the XXth century, but often the scope of "information" is misunderstood. The source of confusion is related to the lack of any referential element in the formula. This has the enormous advantage to "exorcize meaning". Indeed, meaning, or functionality or even a simple form of referentiality is an unavoidable source of problems, whose potential solutions are related to philosophical interpretations. However, it is not less true that human language or biological systems use information in a meaningful way, thereby demanding a richer form of information.

In this dissertation a measure of information accounting for a simple referential parameter is explored [30]. This definition does not try to overcome the philosophical problem concerning the meaning of information in biology. Instead, it properly defines, in terms of information theory, several measures of referentiality conservation proposed in the past [61, 69, 87].

2.1.5 Channel Capacity

A key concept of information theory intimately related to mutual information is the so-called "channel capacity", $C(\Lambda)$, which, roughly speaking, is the maximum amount of bits that can be reliably processed by the system, namely:

$$C(\Lambda) = \max_{p(x)} I(X : Y). \tag{2.17}$$

Channel capacity is an intrinsic feature of the channel; as the Fundamental Theorem of Information Theory [5, 37, 95] states, it is possible to send any message of R bits through the channel with an arbitrary small probability of error if:

$$R < C(\Lambda), \tag{2.18}$$

otherwise, the probability of errors in transmission is no longer negligible. One should not confuse the statements concerning the capacity of the channel with the fact that given a random variable with associated probability distribution $p(x)$, we have:

$$\max I(X:Y) = H(X) = H(Y) \tag{2.19}$$

(provided that $C(\Lambda) > H(X)$). In those cases, we refer to the channel as noiseless. On the contrary, if $H(X|Y) > 0$, there is dissipation of information, and we refer to the channel as noisy. The scenario where $H(X|Y) = H(X)$ is the limit case where all information is destroyed by the noise of the channel.

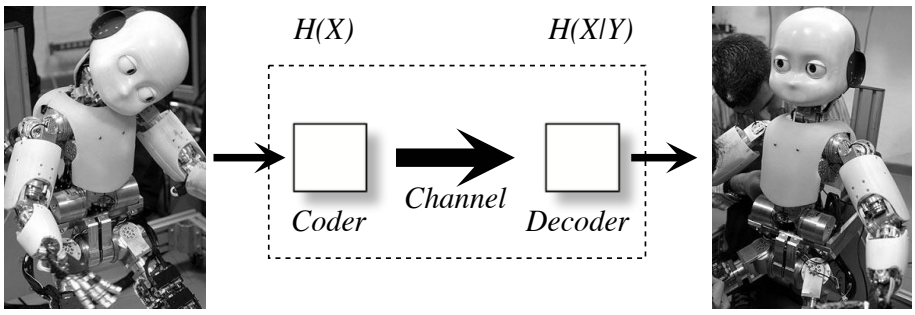


Figure 2.2: A communication system. An external input (objects surrounding the agent or actions taken by other parts of the environment) is coded in some way in signals which, in turn, are sent through the channel. The decoder receives these signals, maybe with errors due to noise and further decodes them into an external output. The central square (dashed lines) represents the essential ingredients required to define information. This highlights the observation that no referential value is actually related to the definition of information (see text).

Figure (2.2) integrates all the elements defining a communication system at the theoretical level. The coder sends in such a way that

its behavior can be described by the random variable X. The entropy associated to X, $H(X)$ is interpreted as the information richness of the source. The channel, Λ, has a bounded capacity, $C(\Lambda)$. Finally, the dissipation of information of the source is evaluated from the random variable Y, accounting for the behavior of the signals once they crossed the channel. Such a destruction of information is evaluated by the conditional entropy $H(X|Y)$.

2.1.6 Channel Capacity and generative codes

Channel capacity plays a central role in the emergence of complex information systems, for it defines an upper bound of reliability. It is also a key theoretical concept when studying the evolution of coding in natural systems. Beyond this upper bound, the presence of noise in unavoidable and, although noise con be sometimes the source of interesting phenomena [59] it can also lead the so-called error catastrophe. The error catastrophe was predicted by M. Eigen, J. McCaskill and P. Schuster [39] and can be informally stated as follows: There is an upper bound to the mutation rate displayed by a population of replicating sequences beyond which information cannot longer be sustained.

The above problem can be directly mapped into a problem of channel capacity: The channel capacity is the maximum size of the sequence that overcomes the problem of the error catastrophe. The case of early replicators and how it was theoretically solved is particularly interesting. The main idea is that, instead of creating longer and longer sequences subject to high mutation rates -high noise levels- we can overcome the problem of conservation of the information by creating small sequences that cooperate thereby forming bigger structures [40]. In a word, to enable combinatorics among existing signals, instead of creating new signals. As will be shown below, a generative algorithm, as the one proposed in [45] and actually overcomes the problem of noise, thereby generating a code with an unbounded

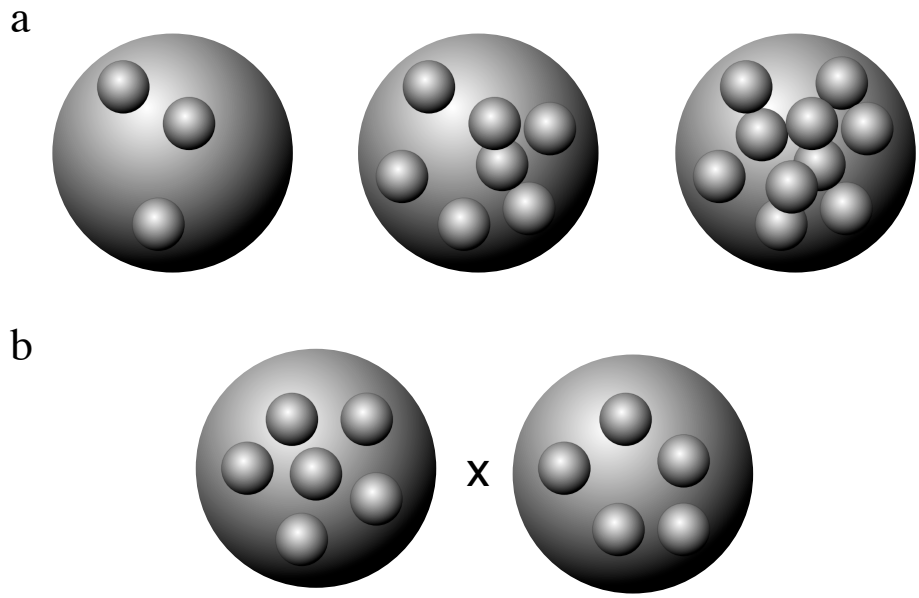


Figure 2.3: a) Let the big circle be the volume of the abstract space where points are possible signals. Red balls represent the region of the space occupied by the signal represented by the point located at the center: If some other signal is represented by a point inside the sphere, the probability of error is non-zero and these two signals can be confused with non-vanishing probability. Above center we show the space saturated of signals i.e., there is a distribution of points by which their associated spheres do not intersect. However, in the upper right corner, we observe that, above a give threshold in the number of signals, intersection between spheres is unavoidable, which can be informally related to the channel capacity. b) If, instead of increasing the number of signals we combine existing signals - which have a configuration where the error probability is zero- the space available for the signals is the cartesian product of two spaces identical to the original one, thus exponentially increasing the number of possible signals and linearly increasing the amount of information that can be carried out by the existing signals.

informational capacity[6].

The roots of the selective advantage of a code which uses combinatorics -or more sophisticated concatenation procedures- of signals stems from the Channel Coding Theorem outlined above [37]. This theorem can be interpreted in the following way: Instead of supposing a rate R of bits per time unit, we can interpret that our set of signals $\mathcal{X}$ is defined in an arbitrary space $\mathbb{U} \subset \mathbb{R}$. Every signal $s_i \in \mathcal{X}$ is depicted by a coordinate point $\mathbf{x}_i \in \mathbb{U}$. The finite resolution of any measurement system and the fluctuations due to thermal noise generates a spherical cloud ν_i around $\mathbf{x}_i$ of a given radius ρ_i. All pair of signals by which the intersection is non vanishing are susceptible to be confused. In formal terms,

$$(\forall s_i, s_j \in \mathcal{X}) \quad \int_{\nu_i} dv = \int_{\nu_j} dv = v.$$

(Notice that we talk generally volume, but in this case we are computing the length of a given interval contained in $\mathbb{U}$ and centered in $\mathbf{x}_i$) As we stated above, the main idea is that two signals cannot be confused due to the presence of noise if the spatial distributions of signals across the interval $\mathbb{U}$ is such that:

$$\int_{\nu_i \cap \nu_j} dv = 0.$$

The volume v will thus determine the resolution of the system. If the channel is noisy, the volume of a given signal will eventually increase until the complete destruction of any information, which implies that

$$(\forall s_i, s_j \in \mathcal{X}) \quad \int_{\nu_i} dv = \int_{\nu_i \cap s_j} dv = \int_{\mathbb{U}} dv. \tag{2.20}$$

[6]This particular derivation cannot be considered a new result, but is the way I found to formally express something that I've found in many places but for which I could not find a clear, well written derivation. I added this derivation here due to the central importance of the selective advantage of complex information systems in biology.

In this system, without further information, to receive the coordinates of a given point $\overrightarrow{x} \in \mathbb{U}$ tells us nothing about the sent signal, since all signals are confused in a gibberish. Alternatively, if the channel is noiseless and accepts to process more and more information, we can find a code containing an arbitrary number of codewords such that:

$$(\forall s_i, s_j \in \mathcal{X}) \quad \int_{\nu_i} dv = \int_{\nu_i \cap \nu_j} dv \to 0. \tag{2.21}$$

In this interpretation, $C(\Lambda)$ determines the maximum number of signals having a non-zero volume that can be spread through $\mathbb{U}$ in the following way:

$$\log |\mathcal{S}^*| < C(\Lambda) < \log(|\mathcal{S}^*| + 1), \tag{2.22}$$

being $|S^*|$ defined as:

$$|\mathcal{X}^*| = \max_{|\mathcal{X}|} : \left[(\forall s_i, s_j \in \mathcal{X}) \quad \int_{s_i \cap s_j} dv = 0 \right]. \tag{2.23}$$

In this case, if v is the volume of the clouds associated to the resolution of signals, it is straightforward to obtain $|\mathcal{S}^*|$, namely:

$$|\mathcal{X}^*| = \max_{m \in \mathbb{N}} : \left(m \leq \frac{1}{v} \int_{\mathbb{U}} dv \right), \tag{2.24}$$

which corresponds to the highest number of signal coordinates we can spread in an equidistant way along the interval $\mathbb{U}$ by which condition (2.21) is satisfied for all pairs of signals.

Having defined the maximum number of signals, if we want to avoid any probability of error, how can we face the selective pressure pushing the system to be able to process more and more information? One strategy is to attribute functional meaning to combinations of signals i.e., that, for example the string $s_i s_k s_j$ has a unique, new interpretation. If we introduce combinatorics of, namely n elements to our system, the different coordinates of combinations of signals $\sigma_i \in \mathcal{X}^n$,

the set of relevant elements will be points of a n-dimentional subspace, i.e., the coordinates of the string σ_i will be. $\mathbf{x}_i \in \mathbb{U}^n$. Additionally, it is clear that $v(n) \subset \mathbb{U}^n$. Now finding the configuration maximizing the number of hyperspheres but ensuring non-overlapping is a formidable combinatorial problem, known as sphere packing problem, for which a few is known if $n > 8$. Thus we will consider that our signals are points of an n-dimentional lattice where the nearest neighbors are separated a distance 2ρ. We know that this configuration satisfies (2.21), although we don't know the exact form of the configurations that, yet satisfying eq. (2.21), contain more signals than the proposed lattice. Thus,

$$|(\mathcal{X}^n)^*| \geq |\mathcal{X}^*|^n \tag{2.25}$$

which is the number of elements of our lattice. Thereby, the potential amount of information we are going to be able to process will be:

$$I^*(n) \geq n \log |\mathcal{X}^*| \tag{2.26}$$

Combinatorial procedures lead to strings of finite size and, thus, what we achieve is a translation of the upper bound in the amount of information we are able to process to higher, but finite quantities. This is how information storage increases in the hypercycle model of Eigen and Schuster, to explain the emergence of replicators overcoming the mutation rate while increasing in complexity [39, 40].

Instead of using combinatorial procedures, let us now suppose we use some generative algorithm -an algorithm whose expressive power is equal or higher than a context-free grammar. Let the (now unbounded) set of functional/meaningful elements to be $\phi_i, ..., \phi_k, ...$ Now our strings are members of spaces of arbitrary dimension and, since there is no bound in the length of the string, one can find a string such that,

$$(\forall n)(\exists \phi_k \in \Sigma) : \left[\int_{\nu_k} dv \subset \mathbb{U}^{n+1} \wedge \int_{\nu_k} dv \nsubseteq \mathbb{U}^n \right]$$

Following the reasoning provided above, we can conclude that the amount of information we are able to transmit is unbounded, i.e.

$$(\nexists \mathbf{M} \in \mathbb{N}) : (\mathbf{M} > I^*). \tag{2.27}$$

Thus, a generative algorithm [19, 57] is able to "solve" the problem of the channel having low resolution, having an unbounded informative power. As shown in [35] this is not the end of the story, because, even in the case of having infinite signals, the code must hold additional conditions -see [31, 35] and references therein.

2.2 Computational Complexity

One of the most striking achievements of the second half of the twentieth century mathematicians is the demonstration that all the basic theorems related to the foundations of mathematics can be formalized in a framework analog to information [18, 68]. Moreover, strict connections among concepts of information theory and computation theory -taking the latter as the core of the foundations of modern mathematics- can be found, defining a new branch of mathematics, namely the Algorithmic Information Theory (henceforth AIT). This merging opens a great deal of possibilities and new interpretations of old phenomena. Following this spirit, in this dissertation some concepts are grounded on the AIT, with the aim to take profit of the great unifying power it possesses. The key concept briefly revised here is the the Kolmogorov complexity or Program size complexity[7] and the connections it has with Shannon entropy.

[7]The reader should be aware that there are several different concepts commonly referred as computational complexity, which, even not completely unrelated, they are definitely different. The first refers to the minimal program size needed to describe an object, which is the one described by the Kolmogorov Complexity, being intimately related to the fundamental problems of computability and decidibility. The second one refers to the speed of convergence of a given algorithm, raising the fascinating field of the NP problems that arise in combinatorial algorithms. The third one is related to the level on the Chomsky Hierarchy of a given algorithm. In this section, and throughout the dissertation, we will

2.2.1 Kolmogorov Complexity

Any algorithm can be described in a more or less lengthy way by a (finite) binary string consisting of 0's and 1's. Thereby, any computable[8] object can be described as the output of a program written in a binary code to be executed in an universal Turing machine[9], $\mathcal{T}_u$. What is the minimal length of this binary program? or, in other words, what is its "Kolmogorov complexity"? The exact, general answer for this question is not possible, for it falls into the category of non-computable problems. This apparently undesirable feature could prevent the reader against the use of Kolmogorov complexity as a concept suitable to construct any physical theory. This leads us to the following philosophical problem: Since Kolmogorov complexity is a conceptual precursor of the entropy, the problem diffuses to

use the first definition, uneless the contrary is indicated. See [57] and [82] and references therein.

[8] Any set (finite or infinite) describable through an algorithm is called computable [38].

[9] A Turing machine, $\mathcal{T}$, is an abstract computing machine that consists of a head or cursor and a potentially infinite tape divided into squares each of which contains a symbol. The head of $\mathcal{T}$, after reading the symbol of a particular square, can either (a) write a new symbol on that square (thereby deleting the formerly read symbol) or (b) move leftwards or rightwards to a new square of the tape. The particular operation that $\mathcal{T}$ carries out (writing, leftward movement or right movement) is determined by the symbol that the head reads and the internal state of the machine. Its behavior is governed by a finite, consistent and deterministic set of transition formulae called quintuples, which, in the general case, display the following form -For the sake of clarity, and following e.g. [9] we define $\mathcal{T}$ by not explicitly formalizing the movement of the head and the accepting states:

$$\langle Q, \Omega, \mathcal{S}, g_{\mathcal{T}}, \sigma_0 \rangle.$$

In this equation, Q is a finite set of states, being $\sigma_0 \in Q$ the initial state. Ω is the input alphabet, $\mathcal{S}$ the output alphabet, $g_{\mathcal{T}}$ the transition function, $g_{\mathcal{T}} : Q \times \Omega \rightarrow Q \times \mathcal{S}$. In general, we admit that either $\Omega \cap \mathcal{S} \neq \varnothing$ or$\Omega \cap \mathcal{S} = \varnothing$. A Universal Turing machine is a Turing machine that can implement any algorithm in some way [38]. By virtue of the Church-Turing thesis, any computable object can be described in terms of Turing machines [38].

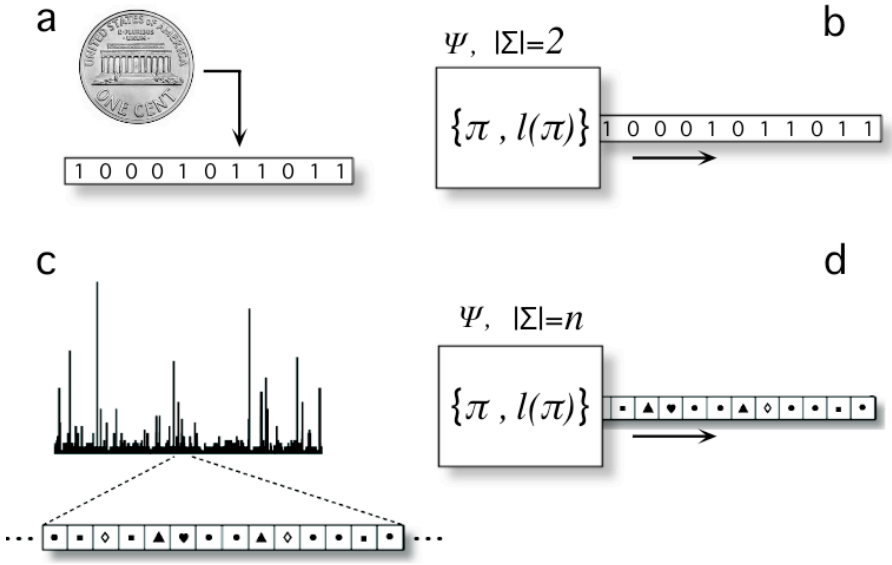


Figure 2.4: Algorithmic complexity $K(\mathbf{x})$ of a stochastic string $\mathbf{x}$, indicating a set of observations made on a system, is measured as the lenght in bits of the minimal program p required to reproduce such string. For a fair coin toss (a) which generates a completely random sequence, the computer Ψ (b) would run a program with a length equal to the lenght of the string (which is an upper bound here). Here the size of the alphabet Σ is two (i. e. $|\Sigma| = 2$) but an arbitrary sequence (c) $\mathbf{y}$ obtained from the successive observations over complex system would not be restricted by the binary description. Instead, a large range of n possible symbols would be used to define our string now. This is coded through a minimal program which, when applied to a computer (d), replicates the $n-$ary original sequence. The length of this minimal program, coded in bits, is the Kolmogorov Complexity of $\mathbf{y}$.

the concept of entropy. And indeed it is so: When we say that " the probability of the event s_i is $p(s_i)$" we are assuming total knowledge of the behavior of the system; moreover, a perfect reliability of our (abstract) measurement technologies. Can we ensure this perfect reliability? Obviously not, for it will imply the existence of a platonic reference [49, 50].

Can we ensure (assuming perfect reliability) that the probability of event s_i is $p(s_i)$ in a finite number of measurements? of course, by the law of large numbers, we cannot. So the uncomputability problem of Kolmogorov complexity is also present in the computation of the entropy of the system, although its use is widely accepted. Jaynes' viewpoint over statistical physics is, in some way, the other side of the coin: Assuming partial knowledge, what is the most likely configuration of the system? Jaynes' subjective viewpoint says nothing about the computation of the entropy in a given system, and the problem of computability still remains. This is, of course, a fascinating problem of the foundations of science, and the interested reader can go through [49, 62].

Let x be a finite binary string binary string and let $\mathcal{T}_u$ be a universal Turing machine. Let $l(x)$ be the length of the string x. Let $\mathcal{T}_u(p)$ denote the output of the computer $\mathcal{T}_u$ when presented with a program p. The Kolmogorov Complexity $K_{\mathcal{T}_u}(x)$ of a string x with respect to a universal computer $\mathcal{T}_u$ is defined as:

$$K_{\mathcal{T}_u}(x) = \min_{p:\mathcal{T}_u(p)=x} l(p). \tag{2.28}$$

Interestingly enough, this quantity is computer independent up to an additive constant, thus we will leave the subindex to refer to it, leading us to write K alone. Now suppose that a computer is fed a random program. It is analogous to suppose that we define the input string x from $l(x)$ coin tosses, thus describing a random string made of $l(x)$ 0's and 1's. What is the probability that such a Universal computer will print the desired object as the output of the computation? The answer comes from the definition of the Universal probability, P_u, which tells us the probability that a given $\mathcal{T}_u$ will print the desired

output string considering it is fed with random sequences of length $l(x)$, namely:

$$P_u(x) = 2^{-K(x)}. \tag{2.29}$$

If we assume that the computer already knows the length of x, then we can define the conditional Kolmogorov complexity, as:

$$K(x|l(x)) = \min_{p:\mathcal{T}_u(p,l(x))=x} l(p) \tag{2.30}$$

This is the shortest description length if the computer $\mathcal{T}_u$ has the length available to it. In figure (3.1) we detail the definition of Kolmogorov Complexity for abstract stochastic objects, which are the kind of objects we are interested in.

2.2.2 Defining complexity

Kolmogorov complexity can be considered a conceptual precursor of Shannon's entropy. Furthermore, it solves a problem of consistency within Shannon's proposal, namely, the problem of absolute information content of a given object. Notice that Shannon entropy -see eq. (2.1)- is computed by taking into account the relative abundance of a set of events in a given stochastic process, and it is strictly zero if there is only one event. Furthermore, K is the minimal number of bits to obtain a precise and unambiguous description of a mathematical object, in the most general way, whereas Shannon entropy is the minimal description of the behavior of a stochastic object, a specific kind of mathematical object -see eq. (2.4) and figure (3.1). Not surprisingly, they are intimately connected.

In mathematical terms, a sequence of observations (obtained from a given system) whose outcome is probabilistic is a stochastic object. By definition, the Kolmogorov Complexity of a stochastic object described by a binary string $\mathbf{x} = x_1, .., x_m$ of length m, satisfies the following requirement [50]:

$$\lim_{m\to\infty} \frac{K(\mathbf{x}|m)}{m} = \mu \quad \in (0, 1]. \tag{2.31}$$

In other words, the binary representation of a stochastic object is linearly compressible. The case where $\mu = 1$ refers to a completely random object, and the string is called incompressible. As an example, let us consider a Bernouilli process, described by a binary random variable X such that $\mathbb{P}(X = 1) = \theta$ and thus $\mathbb{P}(X = 0) = 1 - \theta$, [37]. Suppose we perform m observations, thereby generating the sequence of independent, identically distributed random variables

$$X_1, ..., X_m, \tag{2.32}$$

which is, in this case, a sequence of 1's and 0's. The Kolmogorov Complexity of the string generated by a sequence of m observations over such a stochastic system, $K(X_1, .., X_m)$, satisfies the following scaling relation [50]:

$$\lim_{m\to\infty} \frac{K(X_1, ..., X_m)}{m} = \mu; \;\; \mu \in (0, 1]. \tag{2.33}$$

In this case it is straightforward to identify [37]

$$\mu = H(\theta), \tag{2.34}$$

where $H(\theta)$ is the uncertainty associated to the Bernouilli $\sim \theta$ process, i.e., its Shannon entropy -see eq. (2.1):

$$H(\theta) = -\theta \log \theta - (1 - \theta) \log(1 - \theta) \tag{2.35}$$

The average Kolmogorov Complexity is tied to the uncertainty in predicting, from a given row, the value of the next row, either 1 or 0 -see fig . Notice that the most uncertain case is obtained for $\theta = 1/2$, leading to $H(\theta) = 1$, according to the definition of randomness provided above.

We can generalize the concept of random sequence for non binary strings, whose elements belong to a given set $\Sigma = \{s_1, .., s_n\}$, being $|\Sigma| = n$ -see figure (3.1c,d). This is the case of a dice, for example, whose set of outcomes is $\Sigma_{dice} = \{1, 2, 3, 4, 5, 6\}$. Accordingly,

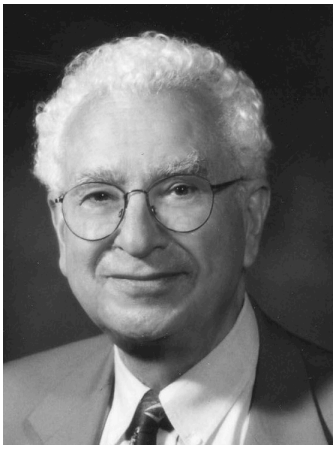
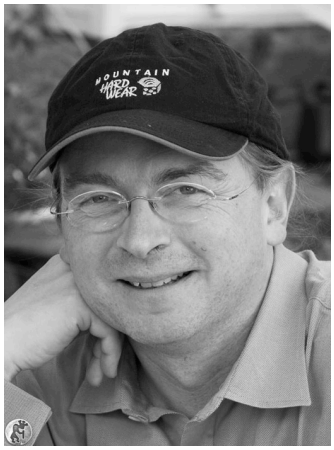



Figure 2.5: The problem of how to properly define complexity based on information theoretical approximations was addressed by physicists Murray Gell-mann (left) and Seth Lloyd, working together at the Santa Fe Institute. They defined the concept of effective complexity in terms of the length of a highly compressed description of its regularities.

the successive observations of our stochastic system are depicted by a sequence of independent, identically distributed random variables $X_1, ..., X_m$ taking values over the set Σ and following a given probability distribution p. By virtue of the interpretation of entropy as the average number of bits needed to properly describe a random variable provided in the above section -see eq. (2.4) The average minimum length will correspond to the minimum length of the code, which is, by definition, the Kolmogorov complexity. Thus we obtain the following equality:

$$\begin{aligned} \lim_{m\to\infty} \frac{K(X_1, ..., X_m|m)}{m} &= \sum_{i\leq n} p(s_i) l^*(s_i) \\ &= H(X) + \mathcal{O}(1). \end{aligned} \tag{2.36}$$

Eq. (2.36) has an annoying interpretation, because relates uncertainty with complexity, in other words, randomness and complexity. It is true that the more random the system, the larger the program to exactly describe it. The identification of program length and complexity is clear and rigorous and, thus, Kolmogorov complexity provided a solid in the foundation of Information theory. However, it is also true that it does not entirely grasp the intuitive idea of Complexity. Indeed, a complex system is not free of regularities, as it happens with random systems, but its regularities are complicated and entangled.

With the aim of solving this philosophical paradox, Murray Gellmann and Seth Lloyd -see figure (2.5)- proposed a measure, the "effective complexity" based on the most general interpretation of Kolmogorov Complexity -which includes all kind of mathematical objects, either stochastic or not [47]. Roughly speaking, effective complexity $\epsilon(\mathbf{x})$ is the length in bits of the shortest program which, when applied to a Turing machine, generates, as the output, the description of the regularities of the system x. In a nutshell, it is the length of a compact description of the identified regularities of a given entity. Therefore, a random object is effectively simple, and so is a regular object. But a system having complicated -non random- interrelations among its parts is effectively complex.

Chapter 3

RESULTS AND DISCUSSION

> In our scientific inquiry, we have to distinguish between ”problems” and ”mysteries”.
>
> Noam Chomsky

The four questions presented above largely shape the range and content of the problems described here. In this chapter we summarize the main results obtained through the research done through the PhD research process, trying to properly put them into the context as defined by our Four Questions. As it happens with any problem on evolved complex systems, we have always to consider the evolutionary scenario where changes and transitions have taken place, shaping complexity.

Within the context of language, its historical development has been the outcome of an evolutionary play within an ecological theatre, in the words of ecologist E. Hutchinson. It deals with several levels of organization and thus the play itself is a multilevel one. Language and its evolutionary dynamics can be seen in terms of how brain structures have changed (or co-evolved) with early communication skills. It can also be seen as a process of channel optimization much in the way of an engineering problem. But it can also be ap-

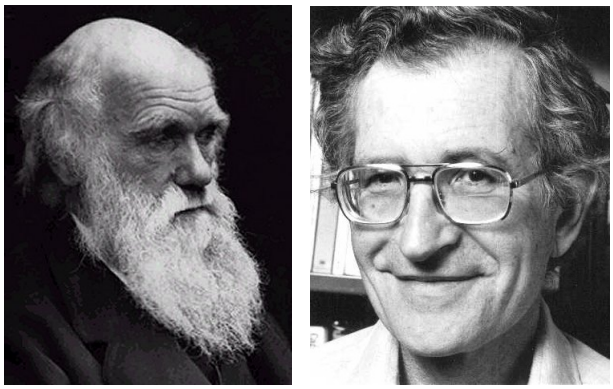


Figure 3.1: Language structure and its evolution have been analyzed by many scholars over the centuries. Charles darwin compared languages with living species, which are able to change, split and get extinct. Such comparison, as shown by modern genomics, was much more accurate that Darwin himself would have suspected. Noam Chomsky (right) one of the fathers of modern linguistics, speculated about the presence of some ”hardwired” readiness for acquiring language that would have been shaped in our brains through natural selection.

proached from an ecological perspective or even in rather abstract terms. The later is particularly relevant here, since human language is characterized by a virtually infinite capacity of building (and not just combining) grammatical structures in meaningful ways. One of our goals here is to address this difficult problem using quantitative approaches.

The species-level view of language in terms of standard biology was first raised by Charles Darwin, who mentioned them a number of times in writings and letters. In The Descent of Man, Darwin explicitly says:

> The formation of different languages and of distinct species, and the proofs that both have been developed through a

gradual process, are curiously parallel.

This ecological and evolutionary, species-level picture of language has been deeply analyzed by many authors -for a recent review, see [97]. In that particular context, language dynamics can be studied at several scales, from word inventories and their diffusion to grammatical change. Both continuous and rapid (punctuated) patterns have been observed, illustrating the complex nature of the tempo and mode of language change. As we already mentioned in a previous chapter, the interpretation of these results proposed by linguist Noam Chomsky concerning the evolutionary origins of a ”language organ” generated a very passionate debate that is still alive today [20].

We mention again this debate since a large number of problems addressed here touch deeply the core of the problem. On the one hand, we want to face the question of the emergence of recursive structures and how to properly formalize it. This requires an adequate analysis of the available evidence from acquisition databases. On the other hand, we also address one of the key problems that has failed to be correctly defined in the treatment of language and communication in terms of information theory: meaning or, in its weaker version, referentiality. Without an appropriate definition of how coding and decoding is made while preserving meaning, the selective value of communication described by any theory of language complexity will be, to the least, questionable. Finally, let us point out that Zipf's law, although considered as non-relevant for human language by Chomsky himself[1] [81], might actually be a crucial piece for understanding and characterizing the unbounded nature of human language. In other words, Zipf's law is not only relevant in order to understand and characterize complex forms of communication. It might also be a signature of the critical requirement for unbounded complexity. The consequences, we believe, go far beyond

[1]The criticism of Miller and Chomsky was focused on the non-relevance of statistical patterns to study the generative rules of language. In this way, Zipf's law was specifically cited as non-relevant because is a paradigmatic example of statistical pattern of human language -see chapter 1.

human language and might be also relevant within the context of general evolutionary dynamics.

Along with a multiplicity of scales of complexity, language itself has been treated in multiple forms within theoretical approaches to communication. This includes computational linguistic methods and information-theoretic tools. Here information theory and its computational implications (as presented in the previous chapter) defines the hardcore of our mathematical approach, completed by the use of complex networks theory. These methods are used in the analysis of real data and in defining our general, theoretical concepts.

3.1 The Ontogeny of Syntax

Among the most remarkable patterns of language change, its ontogeny stands has a very special status. Infants learn words and some rough, necessarily incomplete set of grammatical rules in a rather non-conscious way. And yet, in spite of the fact that only a small fraction of the potential set of samples that can be considered a good covering of the representative rule space, children experience a rapid mastering of syntax around the two years of age. Not only this: the process itself by which syntax is somehow mastered is highly non-linear. After a sequence of babbling, one-word phase and two-word phase, a transition towards complex language takes place. In a nutshell, we could say that the child jumps from two to infinity.

The analysis of syntactic networks belonging to the first acquisition stage revealed many interesting features. At the methodological level, it is important to highlight the crucial fact that the network approach, with all its limitations, enables us to obtain quantitative estimators of linguistic complexity taking into account the role of the combinatorial ingredient. Therefore, the so-called syntactic spurt -i.e., the abrupt emergence of complex syntactic structures beyond the two words stage- can be identified from quantitative grounds. This section, thus, provides tentative answers for Question One: What are

the empirical patterns we can observe along the evolution of a code displaying an increasing in generative power and complexity?

3.1.1 Patterns

We studied the complexity patterns of the networks obtained from two individuals. These two individuals, known as "Peter" and "Carl" in the CHILDES Database [14, 15] have been studied within the time window that goes from the age of around 21 months to to the age of 27 months [27, 33, 36]. The general trend of the quantitative estimators of complexity through the syntactic networks obtained along the acquisition process show clear increase with time with a central region -around the age of two- in which such increase is sharper. Moreover, the networks of the two studied individuals display a very special behavior: From the very beginning, even when most of the words appear isolated, almost all the words participating in some syntactic relation appear in the same cluster. In other words, syntactically active words belong to the same connected component. We can thus advance that this rules out any hypothesis related to percolation to explain the emergence of syntax. This connected component grows in absolute and relative size and, at the end of the process, isolated words seldom appear. Additionally, another feature has to be highlighted: The networks obtained from the beginning of the process display a clearly defined tree structure, whereas beyond the syntactic spurt -clearly identified at the age of two- networks display a scale free behavior with non vanishing levels of clustering and with increasing connectivities -see figure (3.2)- and chromatic numbers -see figure (3.3)-, which is a clear deviation from the tree behavior.

If we go to a more detailed level, we have to distinguish between the data obtained from the two studied individuals. Indeed, whereas the first studied subject -Peter- displays a well shaped curve of complexity indicators, having a more or less stationary regime at the beginning, and then followed by a sudden jump at the age of two, the quantitative observables of the second individual -Carl- are not

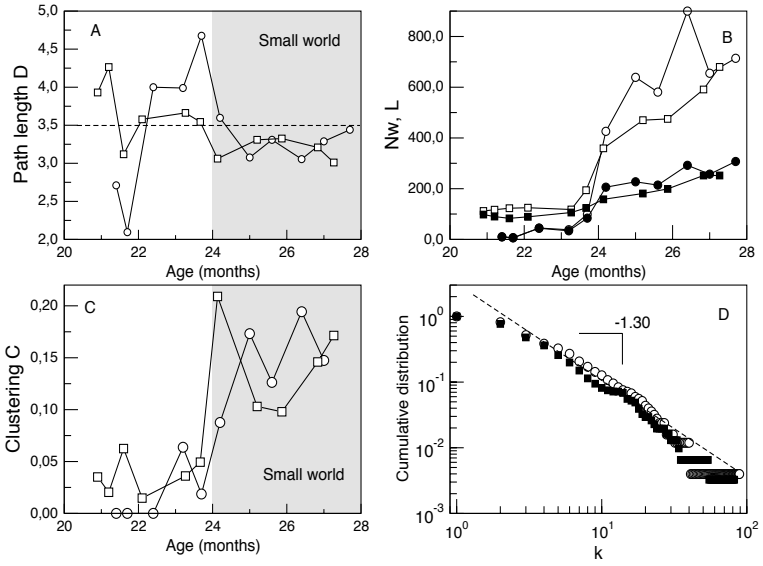


Figure 3.2: Changes in the structure of syntax nets as obtained by using complex networks approaches. Here we show: the average path length (a) the number of words and links (b) the clustering (c) and (d) the final degree distribution after the transition. As shown in (a) and (c) a small world pattern emerges at the transition point.

so well shaped, although they define a clearly growing trend. If we look at the explicit form of linguistic productions, we find an interesting divergence between the two individuals from the very beginning, which could account for these slightly different behaviors: Functional particles -i.e., determiners, prepositions- and inflectional morphology -such as the "s" of the plural- are completely absent in the early studied Peter's corpora and emerge just at the point where complexity indicators display a jump. On the contrary, Carl's corpora display functional particles from the very beginning. Consistently, the first networks of Peter's conversations are trees, which are bipartite by definition -see figure (3.3). Beyond the emergence of functional par-

ticles the net is no longer bipartite and displays chromatic numbers higher than 2. We finally emphasize that the syntactic nature of hubs also changes with the syntactic spurt: whereas at the beginning hubs are semantically degenerated items, such as "it", beyond the shift in the complexity of the nets hubs are functional particles, such as "a" or "the", and the whole net is organized around them.

As we pointed out above, no (known) network process have been able to reproduce the observed behavior. Indeed, the fundamental process underlying network evolution is a computational one -the generation of syntactically well formed sentences- and, therefore, network's behavior is a statistical picture of what is actually working. Consistently, instead of a typical network growing algorithm, we evaluated the complexity indicators of the networks built from a random sentence generator which reproduced the statistics of child's productions -structure length distribution or frequency of words- but where no syntax was at work[36]. Interestingly, such a model was able to reproduce several trends -such as the mean connectivity or the size of the giant connected component- and identifies the syntactic spurt, but fails when trying to reproduce key complexity indicators involving correlations, such as motif structure [36] or the chromatic index [33].

3.1.2 Origins of the dynamical pattern

Studied data suggests that, at the age of two, something new emerges and completely changes the structure and organization of the net. This innovation seems to be related to the functional particles and inflectional morphology, and is synchronous with the end of the two-words stage. Indeed, one could expect a three-words stage and so-on, which would be reflected in a softer evolution of complexity parameters, but such a period has not been reported in the literature. Therefore, we quantitatively observed the change from the two words stage to complex syntax, where the latter has no restrictions in the potential length of structures. Furthermore, from the very beginning,

syntactically active words are connected and, as we discussed above, the changes in network's organization strongly point to changes in the generative mechanisms of sentence production.

This approach enabled us to explore the emergence of a complex, natural code. The main observation we extract is that such a code seems to emerge, not as the result of a gradual process, but as the result of qualitative changes, which, in turn, imply qualitative shifts in the complexity of the outcome of the code. At the linguistic level, the qualitative shift we observe in data strongly points to some cognitive endowment which, triggered at some point of the language acquisition process -having a special role the continuous stimulation provided by adult speech-, results in the emergence of complex syntax. This would reinforce the chomskyan view that some predefined abilities underlie the process of syntax acquisition.

What is this innovation? In the following section we outline a proposal where a fundamental operation, "merge" is set up at the center of syntactic theory. We will argue that syntactic acquisition is -among many other processes- the picture of the emergence of such an abstract operation, which enables to jump from the two-words stage to a stage where syntactic structure have an unbounded number of words, i.e., from 2 to ∞.

3.2 Generative mechanisms: Merge

Merge has been postulated as the abstract innovation that enabled humans to develop complex language [54]. In this dissertation, by identifying merge as set union [45], we propose a rigorous backbone for the generative theory underlying human syntax, setting up merge at the core [28, 45]. Furthermore, the theoretical skeleton we propose enabled us to generate a family of grammars depending on different parameters which we named "nesting grammars", since the structures belonging to them are the set-theoretical constructs named "nests" [66, 101]. This section tries to provide an affirmative answer to Ques-

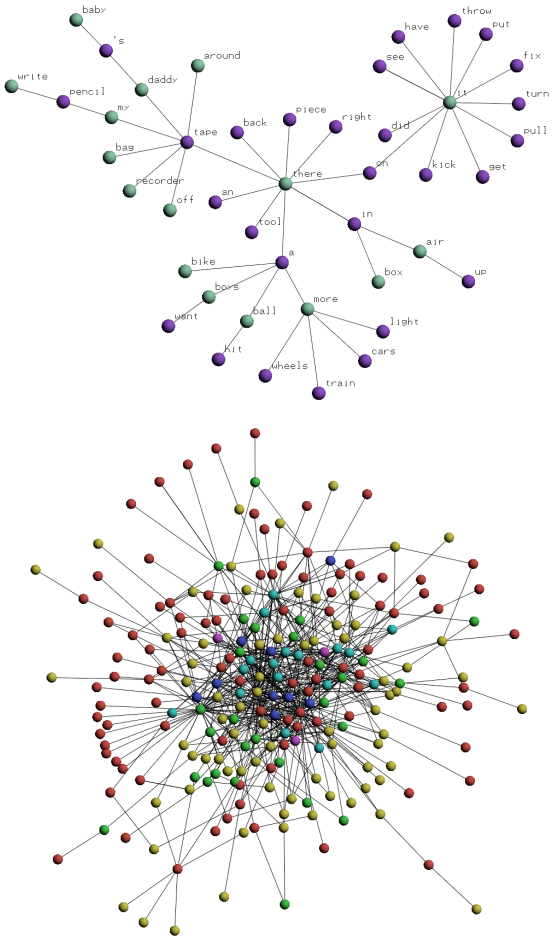


Figure 3.3: A possible minimal chromatic configuration for two networks belonging (above) to the two words stage and (below) beyond the syntactic spurt. The so-called chromatic number of a net is the minimal number of different colors needed to paint the nodes in such a way that no node has any neighbor sharing the same color, and it is a very powerful indicator of network complexity. Interestingly, the two words stage defines a bipartite network where only two colors are needed to properly paint the net, suggesting that such a protogrammar is strongly constrained by compatibility relations. Beyond the two words stage, the chromatic number increases abruptly and the underlying grammar is by far more flexible, suggesting a qualitative, more than quantitative, shift on grammar's properties.

tion Two: Can we develop a minimal theory accounting for the abstract process of syntactic structure generation?.

3.2.1 Nesting grammars

As shown in [45], merge operation, understood as set union, leads to nested structures as the proper mathematical objects from which syntactic concepts, such as syntactic constituent, are rigorously defined. Such objects are the fundamental objects of ordered set theory [66, 101] and we claim that this mathematical framework can play a role in syntactic theory close to the one played by certain areas of mathematics in relation to physics. However, we need to be more restrictive, for it is clear that not all nested structures of e.g., lexical items, are well-formed syntactic objects. To this end, we enriched such a set-theoretical backbone with an abstract structure of compatibility relations, thereby obtaining a complex object: An object that we named nesting grammar. In these grammars, compatibility relations play a role which can be metaphorically related to valence values of a given element: They define what are the compatibility properties of a given element in relation to the other elements of the periodic table to build bigger chemical compounds. As in chemistry, stable elements of our language will be the ones which are neutral, i.e., with no features to check [28].

The nesting grammar is a tuple -see [28], for details:

$$G = \langle \mathbf{A}, \{D_1, ..., D_\mu\}, \Phi, \rho, \delta_c \rangle, \tag{3.1}$$

where

1. $\mathbf{A}$ is the alphabet, which is a finite set of singletons containing the element of A and its features. For every element of an alphabet of singletons A we define a tuple of features, thereby generating the set $\mathbf{A}$, by which $(\forall \mathbf{x} \in \mathbf{A})$

$$\mathbf{x} = \langle x, \langle \varphi_1(x), ..., \varphi_k(x) \rangle \rangle, \quad x \in A,$$

where $\varphi_1, ..., \varphi(x)$ are functions $\varphi_i : A \to \mathbb{N}$, which encode the features of $\mathbf{x}$. These elements of $\mathbf{A}$ -which can be roughly identified as the lexical elements with their grammatical properties- are the bricks of syntactic structures.

2. $D_1, ..., D_\mu$ are the nests that can be simultaneously -but with a finite buffer of memory- generated to build to build complex syntactic structures.

3. Φ is the structure of features and compatibility relations among the features.

4. ρ can be either $r, n/r$ or n, and refers to the way the grammar applies the feature checking, either restrictive (r), neutral/restrictive (n/r) or non-restrictive (n). In a restrictive grammar, two elements can be merged if and only if they are compatible. In a neutral/restrictive grammar, such restriction is relaxed in some cases and, in a non-restrictive grammar, the structure of features and compatibility relations only plays a role in defining neutral structures, but it does not restrict any application of merge.

5. δ_c is the upper bound on the distance of constituents that can be internally merged to generate another constituent. It is s a parameter of working memory over the structure.

The finiteness of memory resources has been studied in depth in [28] and revealed as the source of interesting syntactic phenomena, such as successive ciclicity. In this framework, merge is constrained by a set of compatibility relations K among features, the "state" of a syntactic object is obtained by the application of a "checking function" along the derivation and the state of the elements of the language is "neutral" with respect the compatibility relations.

This abstract formalization of current trends of syntactic theory provides, under our viewpoint, a healthy and well defined mathematical framework for this theory. The aim is to endow theoretical syntax

with a mathematical structure where rigorous results can be derived. It also enabled to properly define the levels of study: From the only structural one -the creation of nests- to the more complex where we introduce features and more abstract elements. As a personal observation that may justify the approach taken here, I think that current syntactic theory is a bit vague in determining the scope of the results obtained within linguistic theory. After the work provided in this dissertation, I would honestly say that theoretical syntax would be restricted to structure generation and some abstract structure of compatibility relations. Beyond this, the interferences of interpretative, communicative and pragmatic issues[2] are, under my viewpoint, stronger enough to be lead into a mathematical gibberish. In saying this, I vindicate that, properly defining its scope, a mathematical theory of syntactic generation must be possible. The philosophical problem set up here, where creative forces are at work is one of the hardest problems that linguistics must face defining, maybe, the borders of the success of any scientific inquiry.

3.2.2 Merge and the Acquisition process

The framework developed in this dissertation enables us to characterize the qualitative changes emerging during the syntax acquisition process. Indeed, during the two words stage, syntactic structures are formed by two words, having these two words complementary semantic features, like $\langle \text{verb}, \text{noun} \rangle$. These semantic constraints are applied in a restrictive way, for it is uncommon to find structures like $\langle \text{verb}, \text{adjective} \rangle$, $\langle \text{noun}, \text{noun} \rangle$, etc [91]. Therefore, we can consider that at the two words stage, we have a grammar of the type:

$$G_{2\text{ words}} = \langle \mathbf{Lex}(t), \{D_1\}, \{\langle \varphi_1, K_1 \rangle\}, r, \delta_c \rangle \tag{3.2}$$

[2] I really think that many of these fields of linguistics include an ingredient of creativity and, thus, current efforts to make mathematical theories of them may fail, thereby forcing us to adopt new ways of scientific inquiry.

where δ_c could be either 0 or 1 and $\mathbf{Lex}(t)$ would be the lexicon of a given language at the age t, i.e., the set of all known words of this language when the child is t_1 years old. Furthermore $\langle \varphi_1, K_1 \rangle = \Phi$, and we emphasize that there is only one feature, φ_1 and its associated compatibility relation, K_1, applied in a restricitve way, r.

This grammar generates a finite language, as shown in lemma 3, [28]. The crucial point comes here: we propose that in this stage we observe the emergence of merge which might be triggered by the increase of the working memory, which enables to be able to deal with more complex structures. Therefore we do not expect to see the three words stage, but if merge is at work, we expect the length of structures to jump ”from two to infinity”.

Let us grasp the above outlined intuitive idea: In the two words stage, the kid generates strings of two words following a grammar like the one proposed in equation (3.2), thereby generating a finite collection of ordered pairs:

$$\langle a, b \rangle,$$

where $a, b \in \mathbf{Lex}(t)$. The qualitative change comes here: instead of making larger strings, the system generates nests: ordered pairs whose members might not be elements of the lexicon, but also already formed nests:

$$\langle a, \circ \rangle$$

where $\circ$ is either an element of the lexicon or another nest. Therefore, one can have structures like:

$$\langle a, \langle b, c \rangle \rangle,$$

where $a, b, c \in \mathbf{Lex}(t')$, $t' > t$. But, as soon as the second member can be a nest, there is no reason to consider the above structure qualitatively different from:

$$\langle a_1, \overbrace{...\langle a_k, ...\langle a_j, a_i \rangle ...\rangle ...}^{\text{arbitrarily large}} \rangle,$$

$(a_1, ..., a_k, ..., a_j, a_i \in \mathbf{Lex}(t'))$. It is crucial to emphasize that the abstract change is that the computational system does not care on the nature on the elements to be nested. Thus, the emergence of merge -the jump to an adult grammar- would imply a transition from a finite-state grammar to a grammar having infinite generative power.

Can we go further? So far we characterized the grammar of the two-words stage and we exposed how merge implies a qualitative shift in the generative power of the system. But, is our formalism able to characterize adult grammars? We hope it is, and we propose the following way: The backbone of theoretical syntax for human language would be defined by the following nesting grammar G_{HL}:

$$G_{HL} = \langle \mathbf{Lex}, \{D_1, ..., D_\mu\}, \Phi, n/r, \delta_c \rangle. \tag{3.3}$$

In this mathematical object, the key piece for further scientific research is Φ, the structure of features and compatibility relations. This key ingredient will define the details of the grammar, beyond its structural features. We can conjecture that

$$|\Phi| \approx 2,$$

being the two kind of features the so-called "formal features" (agreement, etc...) and the so-called "semantic features", (thematic relations, etc) [92]. The kind of application for the compatibility relations would be neutral/restrictive (n/r), for it is flexible enough but, at the same time, restricts some applications of merge. According to standard theories of syntax, we state that, in a given structure, when there are no features to check this structure belongs to the language and can be semantically interpreted. Finally, we consider that there are strong reasons to assume that:

$$\delta_c > 1,$$

therefore, internal merge or cyclic movement is expected to occur.

The above defined grammar is, at least, a context free grammar -as shown in [28]-, although we conjecture that it could be a context-sensitive grammar. This last conjecture is based on the possible role played by the non-vanishing working memory (δ_c), but it is still an open problem.

A consistent theory of language must be able to explain its acquisition process [20]. Concerning syntax, this picture is consistent with a minimal explanation of its acquisition process: Merge is a computational ability that belongs to the cognitive endowment that is at work from the very beginning. However, since the working memory is too short in these early stages -and the lexicon must be still acquired-, we first observe one or two word structures and, then, once the working memory allows to deal with three elements, merge is actually able to generate unbounded structures, due to the nest interpretation. The extreme complexity of morphological and grammatical issues present in adult languages would be the result of the predation of this mechanism for communication, and thus, many factors, such as pragmatic, communicative or psychological ones have to be taken into account[3]. Since every kid generates its own solution -strongly constrained by the input received- languages will never be stable entities.

To close now the linguistic discussion, it is worth to highlight a direct prediction of this theoretical construct: The emergence of merge and the further generation of nested structures rules out the need of postulating intermediate protolanguages having less complexity than the ones observed nowadays. Therefore, the emergence of merge mechanism to be used for communication -thereby having a complex and wide cognitive endowment- would imply a sharp transition from isolated signal-based communication system to a complex grammar like the one we know today. This would explain, at the ontogenetic level, why there is no such three or four-word stage and, at the phylogenetic level, why all languages seem to be, quoting Von Humboldt, equally complex when trying to study them.

[3]The existence of regularities at this level of language would be originated from the interaction of these modules and need not be postulated as fundamental

3.2.3 A general framework

The above outlined generative mechanism describes, in a minimal way, how nested structures can be generated through a fundamental operation, which is no other than set union. This minimal requirement and the strong consequences that can be derived from it makes this generative mechanism suitable to be explored in other fields where the emergence of an unbounded complexity is observed, like the genome.

Although it is still a conjecture, formal approaches such as the Turing gas [44] have been proposed proposed by Walter Fontana as a prebiotic scenario or the Eigen and Schuster's combinatorial solution to the error catastrophe [40] could be reformulated from this framework, enabling, thus, to deal with a -possibly- open ended mechanism of structure generation. However, as we shall see in the next section, the presence of strong generative mechanisms is not enough to properly talk about emergence of unbounded complexity or unbounded information systems, as commonly have been identified. We need another ingredient, and Zipf's law will be at the center of such an additional feature to properly talk about unbouboundedness of complexity or information capacity.

3.3 Zipf's Law = unbounded complexity

In this dissertation we present a general mathematical framework to explain the origin of Zipf's law. The novelty of our approach with respect to other approaches is that our proofs are model-free and they are only based on complexity or information theoretic grounds. The wide range of systems that fall into the characterization provided here would explain the ubiquity of such a law in nature [35]. Additionally, the mathematical framework contains the pieces by which the "least effort hypothesis" [108] -conjectured by G. K. Zipf as the origin of such a scaling law in human communication- can be properly stated. Therefore, we provide a rigorous proof that Zipf's conjecture was

correct [31].

We also explore the very special role of such a probability distribution, which defines a border between systems whose complexity is bounded and the systems whose potential complexity is unbounded. Moreover, this statistical pattern is actually an attractor for the systems whose potential complexity is unbounded. Finally, we discuss the relevance of the above discussions to obtain a tentative, abstract definition of open-endedness in evolution. This section, thus, offers a rationale able to Question Three What is the origin of Zipf's law? Can we provide a mathematical argument accounting for its emergence in the framework of information theory? What is the role of Zipf's law when we study the complex systems -like human language?

3.3.1 Zipf's Law as a nonequlibrium attractor

To grasp the generality of the obtained result, let us set up it in an abstract way. Let us suppose an "stochastic object" whose dimensionality is not fixed but instead grows in time. This can be introduced, without any loss of generality, assuming that $\Omega(t)$ is the set where the random variable $X(t)$ -the stochastic object, in our terms- takes values, and that $\Omega(t)$ evolves in the following way:

$$\begin{aligned} \Omega(1) &= \{m_1\}, \\ \Omega(2) &= \{m_1, m_2\}, \\ ... &\quad ... \\ \Omega(n-1) &= \{m_1, ..., m_{n-1}\}, \\ \Omega(n) &= \{m_1, ..., m_{n-1}, m_n\}. \end{aligned}$$

The elements $m_1, ..., m_n$ belonging to the set $\Omega(n)$ are ordered in such a way that, when observing specific realizations of $X(n)$,

$$p_n(m_1) \geq p_n(m_2) \geq ... \geq p_n(m_n),$$

being $p_n(m_i)$ the probability distribution followed by $X(n)$.

From now on, we described a very generic object which could depict the growing of an urban system -where events are cities- or the distribution of wealth in a market economy -where events can be companies. We make an ansatz: Such systems evolve between order and disorder. They are not completely random systems, but they are not perfectly ordered. Thus their growing is channelized by the window defined by order/disorder tensions. This generic feature is introduced in the most possible abstract way. Indeed, we take the definition of stochastic object from algorithmic information theory and we assume that:

$$\lim_{n\to\infty} \frac{K(X(n))}{\log n}) = \mu \in (0,1), \tag{3.4}$$

In the above equation, $\mu = 1$ would be the case of a completely random object which, by the Shannon-Fano theorem for coding and from the equivalence of Kolmogorov Complexity and Entropy for random objects, would display $K(X(n)) \approx H(X(n)) = \log n$. The case by which $\mu = 0$ depicts a system whose amount of disorder is bounded. Now we put evolution explicitly in the system, by means of the "Minimum Discrimination Information Principle" -see section 2.1.2. This principle ensures that the changes between different stages of the system are governed by a kind of "minimum entropy production principle" by which the divergences between two adjacent states (ensembles) of system's evolution is minimal. Therefore, if we add the $MDIP$ to the ansatz provided in equation (3.4) the problem is summarized as follows: Find the probability distribution p_{n+1} such that minimizes:

$$D(p_n||p_{n+1}) \text{ subject to equation (3.4)}, \tag{3.5}$$

being $D(p_n||p_{n+1})$ the Kullback-Leibler divergence between p_n and p_{n+1}. The result we obtained is that, under these very general conditions, Zipf's Law is the only solution at large values of n. It is worth to note that, asymptotically, the value of μ is completely non-relevant.

Therefore, we demonstrated that, for growing systems reaching an intermediate, stable state between order and disorder, Zipf's law would be expected to emerge in the same way that poissonian distributions are expected for pure stochastic processes which are randomly distributed around a mean value. The nice thing is that Zipf's law is observed in non-equilibrium systems and that its emergence is subject to the path dependence imposed by the $MDIP$, which is presented as a variational principle not tied to equilibrium. The range of applicability is, thus, huge. In the next section we explore how this general formalism is applied to communication phenomena. We show that it actually encodes the rigorous version of Zipf's least effort hypothesis, formulated by Zipf himself as the origin of the distribution having his name in communicative phenomena. Therefore, we have actually a proof of Zipf's conjecture.

3.3.2 Zipf's Law in the communicative context

We consider a communication system composed by a coder and a decoder which is not static. Instead, its associated number of signals grows in time and so does the amount of information that can be conveyed from the coder to the decoder. Furthermore, following Zipf's hypothesis, we assume that there is a tension between the efforts made by both agents. Indeed, in terms of code complexity, the coder trend is to be as ambiguous as possible, with the extreme case where it sends only a signal. Alternatively, from the decoder's viewpoint, the most suitable configuration is the one by which every event has a separate and unique associated signal. This tension is solved by properly defining such efforts and imposing that they are balanced.

The coder agent sends information of an environment Ω, i. e. $\Omega = \{m_1, ..., m_n\}$, whose behavior is depicted by the random variable X_Ω. The informative richness of the environment is, thus $H(X_\Omega)$. Every event $m_i \in \Omega$ is coded through a signal $s_i \in \mathcal{S}$, where $\mathcal{S} = \{s_1, ..., s_m\}$ is the set of signals, whose ordering is related to their probability of

appearance, q:

$$q(s_1) \geq q(s_2) \geq ... \geq q(s_m).$$

The richness of the coding performed by such an agent is the entropy of the random variable, X_s (which follows the probability distribution q) describing the code, $H(X_s)$, and it is identified as the effort of the coder agent. It is straightforward to check that such an interpretation grasps all the properties attributed to coder's effort. The amount of uncertainty in inferring X_Ω faced by the decoder will be directly associated to the effort of the decoder. Consistently, this effort is quantified by $H(X_\Omega|X_s)$. The balance between these two efforts will be:

$$H(X_s) = H(X_\Omega|X_s). \tag{3.6}$$

This kind of equations have been already proposed in the literature [43, 52]. From equation (3.6) we can obtain the following general relation:

$$H(X_\Omega) \geq H(X_s) \geq \frac{1}{2}H(X_\Omega) \tag{3.7}$$

However, we can quickly observe that this condition alone does not provide any clue beyond the emergence of a certain degree of ambiguity in the code. We need to introduce evolution. Indeed, we suppose that the code grows unboundedly and that the environment acts as an infinite reservoir of information. Therefore, we assume that, if $X_\Omega(n)$ is the random variable accounting for the environment when we code n events, its entropy behaves as:

$$\lim_{n\to\infty} \frac{H(X_\Omega(n))}{\log n} = \mu \in (0, 1].$$

Then, it is straightforward that, if $X_s(n)$ accounts for the behavior of the code when Ω has n elements, from equation (3.7)

$$\lim_{n\to\infty} \frac{H(X_s(n))}{\log n} = \nu \in (0, \mu],$$

which is an equation of the same family of equation (3.4), thanks to the equivalence of the entropy and Kolmogorov complexity in stochastic systems. If evolution is guided by the $MDIP$, we have proven

that "the rigorous definition of Zipf's least effort hypotheses actually leads to Zipf's law", i.e.,

$$q(s_i) \propto \frac{1}{i}.$$

It is worth to note that the least effort hypothesis leads to Zipf's law only if we impose evolution (and thus, path dependence). The key issues for a communicative system to display Zipf's law are, thus [31]:

- The unbounded informative potential of the code,
- the loss of information resulting from the symmetry condition, depicted in eq. (3.6), and
- evolution, and its associated path dependency, variationally imposed by the application of the $MDIP$ over successive states of the evolution of the system.

3.3.3 Zipf's Law at the edge of Infinity

As we pointed out above, one of the most striking features of Zipf's law is that it acts as an attractor for systems which unboundedly increase in complexity, depicted in a general way by equation (3.4), no matter the value of μ. In this section we want to highlight an important implication of this distribution which we consider a key result of our work, namely its role as the distribution defining the edge between potentially infinite information systems and bounded information systems.

Suppose that we have a system that grows in time like the one shown in section 3.3.1, having, at every stage of the evolution, a random variable $X(n)$ describing its behavior. Let us consider that $n \to \infty$ and, thus, that there is no bound in the number of states that it can achieve. Let p_n be the probability distribution of the ssytem at the stage n of the evolution. If Zipf's law dominates the probability

distribution, i.e., if $\exists n^*$ such that, $(\forall \delta > 0)(\forall n > n^*)$

$$(\exists k < n) : (\forall i)(n \geq i > k) \left(\frac{p_n(m_{k+1})}{p_n(m_k)} < \left(\frac{k}{k+1} \right)^{1+\delta} \right) \qquad (3.8)$$

thus the complexity of the system is bounded, even when the system itself is infinite. On the contrary, if the probabilistic description of the system obeys Zipf's law or some distribution that dominates it, system's complexity is unbounded. In plain words, what equation (3.8) says is that, if the probability distribution decays faster than Zipf's law, then its complexity is bounded. If the probability distribution describing the system is not dominated by Zipf's law its complexity increases unboundedly with the size of the system. This results can be directly derived from the convergence properties of the Riemann-Zeta function [1] -see [35] for details.

This means that Zipf's law is a footprint of infinity (in terms of information capacity or potential complexity) in complex systems. Indeed, a system exhibiting Zipf's law can overcome any complexity threshold. This is crucial in the context of our study: An infinite generative system must be accompanied by the capacity to "use" it displaying a statistical behavior equal or not dominated by Zipf's law. Otherwise, despite the infinite potential of the system, the information conveyed by it will be necessarily bounded. In other words, we need the system to be able to generate signals following, at least, Zipf's law or another flatter probability distribution, in order to allow a transmission of an unbounded amount of information. In the case of human language, thus, Zipf's law would be observed in parallel to the generative mechanisms able to generate an infinite number of possible sentences. However, a word of caution is needed: As far as we know, Zipf's law has been reported for word frequencies. Word generation rules change from language to language and maybe there is no reason to believe that word-generation mechanisms are qualitatively different from the ones we shown in section (3.2). However, it is well accepted that the inventory of words, even astonishing, is finite. The observation of truncated power laws in the statistics of words

in real languages (Cancho and Sole) would be the footprints of this finiteness. However, the signals of our code are not words, but sentences, generated by the unbounded mechanism merge. Therefore, the conclusion that human language is able to convey an unbounded amount of information would be taken if the statistics over sentences, not words follows a Zipf's-like probability distribution with truncation at high values. This implies, of course, an enormous effort of data analysis which is beyond the scope of this dissertation. Our conjecture is that this scaling should be expected to occur, due to the role of creativity. Creativity would push the system to be assymptotically uncomputable, for it is a concept frontally opposed to computability[4]. Issues related to computability and non-computability are discussed in the next section.

3.3.4 Consequences for open-ended evolution

The previous interpretation of Zipf's law opens a new view to very important problem within complex systems evolution, namely the conditions for open-ended evolution [65]. This refers to an evolutionary process where complexity can be constantly created and increased. Such problem has been specially addressed within the context of artificial life, by considering the possibility of building a system able to evolve under artificial conditions and maintain a constant source of "creativity". It is important to mention that this problem is deeply connected to the language complexity studied here, both at the level of structural complexity and generative potential. In this context, early work by a number of researchers, including Stuart Kauffman and Walter Fontana -see figure (3.4) showed that abstract systems (such as set of cross- and autocatalytic molecules) able to display constructive rules of interaction and "polymerization" where able to develop complex hierarchical structures with an internal organization describable in terms of grammars [44, 65].

[4]It is worth to note that the definition of computability explicitly rules out creativity [38].

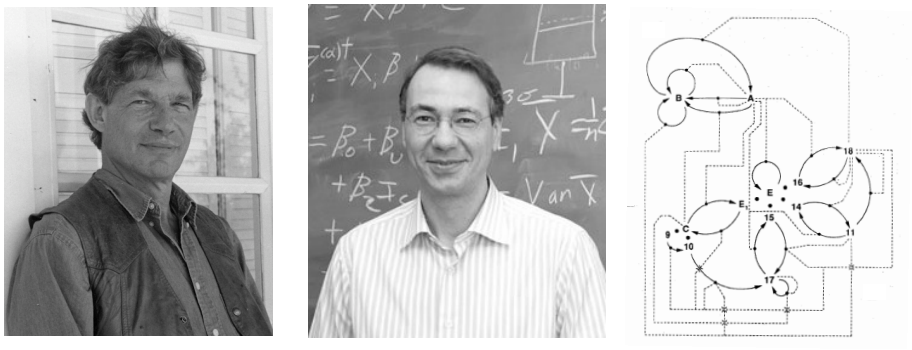


Figure 3.4: Stuart Kauffman (left) and Walter Fontana were among the first scientists pointing towards the problem of open-ended evolution in artificial systems and their relevance for evolution and evolvability. They also showed that complex evolved systems in-silico can display spontaneous emergence of hierachical organization. Such forms of organization can be formally approached in terms of formal languages. The artificially evolved systems (right) typically display feedbacks and multiple levels of dynamical structure.

Our approach is based on the previous results derived from stochastic systems, but its consequences might well go beyond them. Let us now take the computation-theoretic interpretation of equation (3.4). From its definition, we know that the Kolmogorov complexity accounts for the size of the minimal program whose execution in an universal Turing machine generates a complete description of the system. Therefore, equation (3.4) tells us that, for the systems of interest here, the size of the possible algorithm to account for them does not converge. Even worse; since no such convergence occurs, one could expect that the algorithm itself changes. Thus, there is no algorithm to account for the whole process of evolution. If it is the case, we say that such systems are "open ended" in evolutionary terms. Interestingly, we are identifying the concept of computability with the concept of open-endedness: From the definition of universal

probability, we know that the probability of halting for a Universal Turing machine when it is fed with a program of size $K(x)$ is:

$$P_{\mathcal{U}} \approx 2^{-K(x_t)}.$$

If x_t is our system at time t, and its complexity grows in time,

$$P_{\mathcal{U}} \to 0,$$

and therefore, the system is "asymptotically uncomputable". In the case of growing stochastic systems, we will say that such systems are open ended if and only their probabilistic description is provided by a probability distribution that either is the Zipf's law itself, or it is not dominated by it.

This definition of open-endedness in relation to computability issues has the virtue of being both rigorous and that grasps the conceptual requirements intuitively stated. It has however, a problem, when talking about complex systems. As we pointed out in section (2.2.2), Kolmogorov complexity has the annoying consequence that random systems are the most complex ones. And it collides with the intuition we have about complex systems. The concept of effective complexity better encodes what we need: Effective complexity is the size of the minimal program accounting for the regularities of the system. Therefore, we will outline how to properly define open endedness for complex systems in a programmatic way. Let the effective complexity of the system at time t be denoted by $\epsilon(\mathbf{x}_t)$ Then, a system displaying an open ended evolution should be the one by by which

$$(\forall \mathbf{C} \in \mathbb{N})(\exists t > 0) : (\epsilon(\mathbf{x}_t) > \mathbf{C}).$$

Then, we recover the computational interpretation but being more cautious in what kind of complexity we are interested in. Indeed, it is clear that the complete description of the system, including both regularities and random components will be asymptotically non-computable. This rationale is just the starting point, because one has to care about what regularities are, and how complexity is acquired

through evolution. In that way, recent works on the definition of information gain in evolution, centered on the complexity of the evolutionary path [25, 29, 32] rather than the complexity of individuals can provide the clue on how to define regularities.

3.4 The problem of referentiality

Now we change our focus of interest. So far we have been focused on the emergence of complexity through time, with an special focus on the emergence of complex codes, taking, as a case study, human language. Even the emergence of complexity and its information-theoretic issues are at the core of both language emergence and the emergence of biological complexity, we need another ingredient, to account for a complete theoretical picture of human language or complex communication systems, whatever their nature, if the emerge from an evolutionary process and they are not designed. We need referential values, a piece which is not taking into account in the engineer-like formulation of information theory. Such values can be an abstract form of meaning or the functional character of the pieces of the code. It is indeed, an open problem of information theory with strong philosophical implications, since they knock at the door of the concept of meaning. Acknowledging the extreme complexity of the problem, we choose a simple but illustrative way to explore the role of referentiality, thereby providing a starting point to answer question **iv**): Can we quantify, in terms of information theory, the degree of conservation of the referential value in a given communicative exchange?.

To explore this problem in a formal way, we choose a simplified version of Ferdinand de Saussure's [94] "duality of sign" [61, 69]. In this approach, any signal of the code has an associated referential value from the external world. We consider populations of autonomous agents which communicatively interact between them and that their success is evaluated, not only through the quantification

of mutual information, but also quantifying the degree of conservation of the referential value. Such conservation is considered to be a selective advantage. The elements of this general scenario are such autonomous agents, which are defined as a pair of computing devices,

$$A^v \equiv \{\mathbf{P}^v, \mathbf{Q}^v\}, \tag{3.9}$$

where $\mathbf{P}^v$ is the coder module and $\mathbf{Q}^v$ is the decoder module.

3.4.1 No self-consistency paradox

We first focus our attention to a specific problem early raised in the literature focused on the emergence of shared codes in communities of of communicating, autonomous agents [61, 69], the so-called problem of self-consistency. From the very beginning of the studies on the emergence of the communication based on evolutionary game theory, it was pointed out that a global communicative optimum could be reached even the internal configurations of a given agent were completely incompatible, i.e., that an agent could perfectly communicate with other agents but not understand itself [61, 69]. As shown in [34], this "paradox" is only apparent, and an accurate study of the behavior of the nature of the evolution of the communicative exchange rules out the emergence of non self-consistency.

In order to rigorously state the problem, we briefly revise the evolutionary games defined to study the emergence of shared codes in a population of communicating agents [69, 87]. Indeed, let us suppose the following evolutionary game:

1. At $t = 0$, we have an initial population of n agents $A^1, ..., A^n$ as defined in equation (3.9), with randomly defined matrices $\mathbf{P}^i, \mathbf{Q}^i$.

2. At $t = 1$, such agents communicatively interact between them and every communicative interchange is evaluated by

$$F(A^i, A^j) = \frac{1}{2}\left(\mathbf{P}^i \Lambda \mathbf{Q}^j + \mathbf{P}^j \Lambda \mathbf{Q}^i\right). \tag{3.10}$$

-where Λ is the matrix accounting for channel's behavior.

3. The $k < n$ agents having the lowest global fitness, namely the lower value of
$$\sum_{k \neq i} F(A^i, A^k)$$
are removed, and k copies with mutations of the ones having the highest value of the above sum replace the ones that disappeared.

This evolutionary dynamics run until we reach a stable state. F is maximum if:

$$(\mathbf{P}^v, \mathbf{Q}^u, \mathbf{P}^u, \mathbf{Q}^v \in \Pi_{n\times n}) \wedge \left[\mathbf{P}^v = (\mathbf{Q}^u)^T \wedge \mathbf{P}^u = (\mathbf{Q}^v)^T\right], \quad (3.11)$$

where $\Pi_{n\times n}$ is the set of $n \times n$ permutation matrices -see reference [30] for a deeper exposition. We avoid the role of noise for the sake of clarity[5]. The paradox comes here: In the above described situation there is no need, in selective terms, for any special kind of relation between $\mathbf{P}^v$ and $\mathbf{Q}^v$, and the same applies to $\mathbf{P}^u$ and $\mathbf{Q}^u$. Therefore:

$$(\forall A^i)(i \neq k) \max F(A^i, A^k) \nRightarrow \max F(A^k, A^k),$$

which implies that, in spite that the agent A^k has a perfect success in communicating events from the shared world Ω to other agents, it it doesn't understand itself. This phenomena, even having no impact in the success of communicative scenario, has been considered as undesirable and paradoxical. To overcome this inconsistency, it has been proposed that, under a realistic cognitive framework, every agent has a fundamental Lexical matrix, $\mathbf{M}^u$ which defines that fundamental signal meaning associations made by agent A^u [69]. Under this assumption, self-consistency is guaranteed if the agents reach

[5]The general case where the amount of noise makes, for example, that there is not a unique maxima in the rows of $\mathbf{P}$ and $\mathbf{Q}$ deserves special attention and it is not studied here due to its high complexity.

the maximum pay-off. It is worth to not that this assumption implies a qualitative increasing of the mathematical complexity of the properties of the system, as can be seen, for example, in [69].

Actually, the situation by which such a paradoxical situation emerges is improbable, if we accurately study the underlying graph of communicative exchanges among agents within a population, something that has been missing in the previous approaches. Therefore, the assumption of the existence of the lexical matrix would become dispensable, and for simplicity reasons, could be removed from the theoretical models studying the emergence of communication in autonomous agents. We outline the demonstration of this claim in the following lines.

Suppose that we have three agents, namely A^u, A^v, A^w that reached the maximum possible communicative success evaluated over an underlying graph of communicative interactions which is a line, i.e.:

$$A^u \overset{F(A^u,A^v)}{\longleftrightarrow} A^v \overset{F(A^v,A^w)}{\longleftrightarrow} A^w.$$

Under condition (3.11, we extract the following identities:

$$(\mathbf{P}^u = \mathbf{P}^w = (\mathbf{Q}^v)^T) \wedge (\mathbf{Q}^u = \mathbf{Q}^w = (\mathbf{P}^v)^T), \tag{3.12}$$

However, there is no need for $(\mathbf{P}^u)^T = \mathbf{Q}^u$, $(\mathbf{P}^v)^T = \mathbf{Q}^v$ and $(\mathbf{P}^w)^T = \mathbf{Q}^w$ and therefore, under this connectivity pattern, non self-consistent solutions can display maximum success in communicative terms.

Suppose, now, that there is also a communicative interaction between A^u and A^w and that $F(A^u, A^w) = \max F(A^u, A^w)$ -i.e., we have a triangle. Then, one should add the following condition:

$$(\mathbf{P}^u, \mathbf{Q}^w \in \Pi_{n\times n}) \wedge \mathbf{P}^u = (\mathbf{Q}^w)^T \text{ and } (\mathbf{P}^w, \mathbf{Q}^u \in \Pi_{n\times n}) \wedge \mathbf{P}^w = (\mathbf{Q}^u)^T.$$

Now the situation changes radically. Indeed, since now we need $(\mathbf{P}^u = (\mathbf{Q}^w)^T) \wedge (\mathbf{P}^w = (\mathbf{Q}^w)^T)$, by imposing condition (3.12) we have, as the only possible configuration:

$$(\mathbf{P}^u = \mathbf{P}^v = \mathbf{P}^w = (\mathbf{Q}^u)^T = (\mathbf{Q}^v)^T = (\mathbf{Q}^w)^T,$$

which is a self-consistent configuration, since $F(A^v, A^u) = F(A^v, A^v) = \max F$, and the same is true for the remaining two agents. It is straightforward to extend this reasoning to any graph configuration consisting in a cycle having an odd number of agents.

Therefore, we demonstrated that if we reached the maximum of communicative success and the topology of the underlying graph of communicative interactions has at least one odd cycle the only possible configurations are self-consistent. It is well known that in any random graph odd cycles emerge with probability $p \to 1$ even at low values of size and connectivity [16]. Thus, in an evolutionary scenario, the paradox of non self-consistency is absolutely unlikely to occur.

3.4.2 How to conserve referentiality

In the previous section we have shown how a specific paradox that jeopardized the simplicity of the approach was actually not a problem, since we demonstrated that its emergence is ruled out by evolution. This theoretical hallmark is based on a function accounting for the conservation of referentiality, F, defined in equation (3.10) [69], [87]. This is the fundamental piece, and, in plain words, it is a counter of how many words, on average, conserve the referential value after the whole process of coding by a given agent A^v and decoding by a given agent A^u, $u \neq v$. This is an statistical approach but, from an information theoretic viewpoint, is unsatisfactory, for it is clear that the number of signals of a given code is not enough to describe its informative potential. One of the objectives of this dissertation has been to derive an information-theoretic functional able to account for the conservation of referentiality when studying communicative interactions of autonomous agents in a shared world. The obtained functional is the so called consistent information and is obtained by weighting mutual information with the referential parameter, σ, which accounts for the "ratio of bits obtained from the observation of consistent input-output pairs against the observation

of all possible input-output pairs [30].

The first observation that, since the coding and decoding modules of a given agent are depicted by different, a priori non-related matrices, in general

$$I(A^v \to A^u) \neq I(A^u \to A^v), \tag{3.13}$$

i.e., the mutual information between agents in general depend on which agent acts as the coder and which agent acts as a decoder. From classical information theory, mutual information is obtained by exploring the behavior of input/output pairs, averaging the logarithm of the relation among the actual probability to find a given pair and the one expected by chance. Now we are interested on how many pairs are consistently referentiated at the end of the process. We evaluate such conservation of referentiality through the definition of the referential parameter of the communicative exchange when A^v acts as the coder and A^u acts as a decoder, written as:

$$\sigma_{A^v \to A^u}.$$

This referential parameter is obtained by averaging the fraction of information we can extract by observing consistent pairs against the whole information we can obtain by looking at all possible ones. Therefore, the amount of Consistent Information, $\mathcal{I}(A^v \to A^u)$, is obtained by weighting the overall mutual information with the referential parameter:

$$\mathcal{I}(A^v \to A^u) = I(A^v \to A^u)\sigma_{A^v \to A^u}. \tag{3.14}$$

And the average of consistent information among two agents, $\mathcal{F}(A^v, A^u)$ will be, thus:

$$\mathcal{F}(A^v, A^u) \equiv \frac{1}{2}\left(\mathcal{I}(A^v \to A^u) + \mathcal{I}(A^v \to A^u)\right). \tag{3.15}$$

Since $\sigma_{A^v \to A^u} \in [0, 1]$, from the defintion of channel capacity and the symmetry properties of the mutual information, it is straightforward

to show that:

$$\mathcal{F}(A^v, A^u) \leq \langle I(A^v, A^u) \rangle \leq \mathcal{C}(\Lambda). \tag{3.16}$$

Equation (3.16) is the information-theoretic counterpart of equation (3.10) [30]. It encodes the actual amount of bits that can be used in a selective scenario where the content of the message plays a role. What are the consequences of an information-theoretic framework including, even in the simplistic way we did, some kind of referential value? The most striking consequence is that, in the general case of having some kind of noise, either in the channel or in the coding/decoding process,

$$I(A^v \to A^u) > \mathcal{I}(A^v \to A^u). \tag{3.17}$$

Equation (3.17) tells us that when some kind referential value -either meaning or functionality- is assumed, the actual amount of information that can be used is lower than Shannon's mutual information. This has deep consequences when studying natural communication or natural information codes. Indeed, if mutual information has been used in the past to account for natural information-transmission problems, in some cases, the results were wrong. And, as stated by John Hopfield, biology uses information in a meaningful way, thus this result applies to many fundamental problems of biology. This introduces a new and, under our viewpoint, unavoidable piece to the view we had over information transfer problems -either at the genotype/phenotype level, or at the linguistic level- we had in the past [30].

Chapter 4

CONCLUSIONS

Four main questions, proposed in chapter 1, guided the research presented here. In chapter 3 we proposed our tentative answers, how they connect to each other, and how they generate new questions to be solved. Now we expose, in a nutshell, these contributions, making explicit their relation to the papers presented in the compendium.

The first question,

> i) ”What are the empirical patterns we can observe along the evolution of a code displaying an increasing in generative power and complexity?”

has been explored through an experimental work on language acquisition. The main achievement is that, using modern theory of complex networks, we quantitatively identified the different shifts present during the evolution of syntax at the ontogenetic level. This approach has the virtue to include explicitly the combinatorial ingredient within the statistical analysis framework. The obtained results show a clear, both quantitative and qualitative shift in the network structure at the age of two, which fits the previously reported syntactic spurt and that supports the idea that at this age something is triggered in the cognitive apparatus of the kid leading to qualitative changes in the properties of the system of linguistic generation. The papers in which this question is addressed are:

1. Language networks: their structure, function and evolution. Ricard V. Solé, Bernat Corominas-Murtra, Sergi Valverde and Luc Steels Complexity 15(6), 20-26 (2010)

2. The ontogeny of scale-free syntax networks: phase transitions in early language acquisition. Bernat Corominas-Murtra, Sergi Valverde and Ricard V. Solé. Advances in Complex Systems 12, 371-392 (2009)

3. Coloring networks of early child syntax. Bernat Corominas-Murtra, Martí Sànchez Fibla,Sergi Valverde and Ricard V. Solé

4. Network statistics on early English Syntax: Structural criteria. Bernat Corominas-Murtra arXiv-org :0704.3708 (2007) (Supporting material for the above papers)

What is this qualitative innovation in the generative system? Could we describe it in a formal and minimal way? This leads to the second question,

> ii) ”Can we develop a minimal theory accounting for the abstract process of syntactic structure generation?”

We proposed a formal apparatus for syntax based on what is supposed to be the fundamental innovation leading human syntax as we know today, i.e. ”merge” operation. This formal approach gravitates around this biologically-based operation. The generative system proposed here is intended to be the backbone of theoretical syntax, and provides clues for the understanding of the emergence of syntax both at the ontogenetic and phylogenetic level. The papers accounting for this contribution are:

1. Some formal considerations on the generation of hierarchically structured expressions. Jordi Fortuny and Bernat Corominas-Murtra. Catalan Journal of Linguistics 8, 99-111 (2010)

2. The backbone of theoretical syntax, Bernat Corominas-Murtra. (Unpublished)

The proposed generative mechanism is able to generate infinite many expressions. This could imply a huge selective advantage for those individuals having a cognitive endowment complex enough to be able to use it. However, this infinite generative power does not map directly to infinite information capacity. This crucial distinction connects with a very widespread statistical pattern of both natural and artificial systems, Zipf's law, which is the object of study for the next question:

> iii) "What is the origin of Zipf's law? Can we provide a mathematical argument accounting for its emergence in the framework of information theory? What is the role of Zipf's law when we study the complex systems (including human language)?"

We answer these questions by first proposing a mathematical framework to study arbitrary systems growing between order and disorder. We demonstrated that in such systems, by introducing path dependence by means of the Minimum Information Discrimination Principle, Zipf's law is an attractor under a very broad range of scenarios, thereby explaining the ubiquity of such a law in nature.

Such a mathematical framework is demonstrated to be compatible with a rigorous version of the communicative tension proposed by G. K. Zipf to account the origin of Zipf's law. Therefore, we provided a proof that Zipf's law indeed emerges from Zipf's conjecture. Moreover, in addition to the role it has as an attractor of growing systems under some internal tension, Zipf's law actually defines the edge between complexity-bounded systems and the unbounded ones. Therefore, its observation is a footprint of the ability of the generative rules to build arbitrarily complex structures -or to carry an unbounded amount of information. The papers in which this contribution is presented are:

1. Universality of Zipf's law Bernat Corominas-Murtra and Ricard V. Solé Physical Review E 82, 11102 (2010)

2. Emergence of Zipf's Law in the Evolution of Communication
   Bernat Corominas-Murtra, Jordi Fortuny and Ricard V. Solé
   Physical Review E 83, 32767 (2011)

The last question involves, the role of meaning/functionality in natural codes, a problem that has been outside the classical mathematical/engineered view of information theory. This question has been stated as follows:

> iv) Can we quantify, in terms of information theory, the degree of conservation of the referential value in a given communicative exchange?

The first contribution provided here is related to an (apparent) paradox arising in earlier models of the emergence of communication where a simple referential value was assumed for every signal. This is the paradox of non-self consistency, in which an agent communicating perfectly with a given community of agents does not understand itself. In this dissertation, we provide a demonstration that such a paradox is actually unlikely to happen, if we carefully look at the pattern of communicative interactions that took place along the evolutionary process.

The second contribution is more general and theoretical. We built an information-theoretic functional, the "consistent information", able to quantify the degree of conservation of the referential value in a given communicative exchange, thereby providing a positive answer to question iv). The relationship between the signal and its referential value is assumed to be direct and no compositional rules are at work. Even in this simple case -whose simplicity enables us to deal it with mathematical rigor- an important consequence can be derived, having important consequences for natural communication. In the presence of noise, the actual amount of information that can be used -the consistent information- is always lower than the classical Shannon information. This last result may have an impact in the current interpretation of noise in natural systems. These results are detailed in:

1. Network topology and self-consistency in language games. Bernat Corominas-Murtra and Ricard V. Solé Journal of Theoretical Biology 241, 438-441 (2006)

2. Coding and decoding in the evolution of communication: Information richness and referentiality. Bernat Corominas-Murtra, Jordi Fortuny and Ricard V. Solé (Unpublished)

Bibliography

[1] M. Abramowitz and Stegun I. (editors). Handbook of mathematical functions, volume 55 of NBS, Appl. Math. Ser. U.S. Government Printing office, Washington, D.C., 1965.

[2] C. Adami. Introduction to artificial life. Springer, New York, 1998.

[3] M. A. Arbib. Brains, machines and mathematics. McGraw-Hill. New York, 1964.

[4] V. I. Arnold. Mathematical Methods of Classical Mechanics (Graduate Texts in Mathematics). Springer, 2nd edition, 1989.

[5] R. B. Ash. Information Theory. New York. Dover, 1990.

[6] F. Auerbach. Das gesetz der bevölkerungskonzentration. Patermans Geograpische Mittelungen, 59:74–76, 1913.

[7] P. Bak, C. Tang, and K. Wiesenfeld. Self-organized criticality: An explanation of the 1/f noise. Phys. Rev. Lett., 59(4):381–384, Jul 1987.

[8] Y. Bar-Hillel and R. Carnap. Semantic information. The British Journal for the Philosophy of Science, 4:147–157, 1953.

[9] C.H. Bennett. Logical reversibility of computation. IBM journal of Research and Development, 17(6):525–532, 1973.

[10] C.H. Bennett and R. Landauer. The fundamental physical limits of computation. Scientific American, 253(1):38–46, 1985.

[11] D. Bickerton. Language and Species. University of Chicago Press. Chicago, 1990.

[12] D. Bickerton. Adam's Tongue: How Humans Made Language, How Language Made Humans. New York: Hill and Wang, 2009.

[13] A. Blank and S. Solomon. Power laws in cities population, financial markets and internet sites (scaling in systems with a variable number of components). Physica A, 287(1-2):279–288, Nov 2000.

[14] L. Bloom, L. Hood, and P. Lightbown. Imitation in language development if when and why. Cognitive Psychology, (6):380–420, 1974.

[15] L. Bloom, P. Lightbown, and L. Hood. Structure and variation in child language. Monographs of the society for Research in Child Development. Serial 160, (40), 1975.

[16] B. Bollobás. Random Graphs. Cambridge University Press, 2001.

[17] G. J. Chaitin. On the length of programs for computing finite binary sequences. J. ACM, 13:547–569, 1966.

[18] G. J. Chaitin. A theory of program size formally identical to information theory. J. Assoc. Comput. Mach, 22:329–340, 1975.

[19] N. Chomsky. Syntactic Structures. The Hague: Mouton & Co. Paris, 1957.

[20] N. Chomsky. Language and problems of knowledge. MIT Press. Cambridge, Mass, 1988.

[21] N. Chomsky. Cartesian Linguistics: A Chapter in the History of Rationalist Thought. New York: Oxford University Press, 1997.

[22] N. Chomsky. A minimalist program for linguistic theory. In Kenneth L. Hale and S. Jay Keyser, editors, The view from Building 20: Essays in linguistics in honor of Sylvain Bromberger, pages 1–52. Cambridge, MA: MIT Press, 1993.

[23] N. Chomsky. The Minimalist Program. Cambridge, Mass: The MIT Press, 1995.

[24] M. H. Christiansen and Simon Kirby. Language evolution: Consensus and controversies. Trends in Cognitive Sciences, 7(7):300–307, 7 2003.

[25] S. A. Colgate and H. Ziock. A definition of information, the arrow of information, and its relationship to life. Complexity, pages n/a–n/a, 2010.

[26] M. Conrad. Adaptability. Plenum Press, New York, 1983.

[27] B. Corominas-Murtra. Network statistics on early english syntax: Structural criteria. arXiv.org:0704.3708 (Unpublished, supporting information of [30] and [36]), 2007.

[28] Bernat Corominas-Murtra. The backbone of theoretical syntax. Unpublished, 2011.

[29] B. Corominas-Murtra, H. Fellermann, R. Solé, and S. Rasmussen. On the interplay of kinetics, thermodynamics, and information in simple replicating systems. In H. Fellermann et al, editor, Proceedings of the Twelfth International Conference on the Synthesis and Simulation of Living Systems, pages 433–445. MIT Press, 2010.

[30] B. Corominas-Murtra, J. Fortuny, and R. Solé. Coding and decoding in the evolution of communication: Information richness and referentiality. (Unpublished)ArXiv e-print. arXiv:1004.1999v1, 2010.

[31] B. Corominas-Murtra, J., and R. Solé. Emergence of zipf's law in the evolution of communication. Phys. Rev. E, 83(3):036115, 2011.

[32] B. Corominas-Murtra, C. Rodríguez-Caso, J. Goñi, and R. Solé. Topological reversibility and causality in feed-forward networks. New Journal of Physics, 12(11):113051, 2010.

[33] B. Corominas-Murtra, M. Sánchez-Fibla, S. Valverde, and R. Solé. Coloring the networks of early syntax. Unpublished, 2011.

[34] B. Corominas-Murtra and R. Solé. Network topology and self-consistency in language games. Journal of Theoretical Biology, 241(2):438–441, July 2006.

[35] B. Corominas-Murtra and R. Solé. Universality of zipf's law. Phys. Rev. E, 82(1):011102, Jul 2010.

[36] B. Corominas-Murtra, S. Valverde, and R. Solé. The ontogeny of scale-free syntax networks: Phase transitions in early language acquisition. Advances in Complex Systems (ACS), 12(03):371–392, 2009.

[37] T. M. Cover and J. A. Thomas. Elements of Information Theory. John Wiley and Sons. New York, 1991.

[38] M. Davis. Computability and Unsolvability (Mcgraw-Hill Series in Information Processing and Computers.). Dover Publications, 1985.

[39] M. Eigen, J. McCaskill, and P. Schuster. The molecular Quasi-species. Adv. Chem. Phys., 75:149–263, 1989.

[40] M. Eigen and P. Schuster. The Hypercycle: A Principle of Natural Self-Organization. Springer-Verlag, Berlin, 1979.

[41] F. Family and P. Meakin. Kinetics of droplet growth processes: Simulations, theory, and experiments. Phys. Rev. A, 40(7):3836–3854, 1989.

[42] P. Fernàndez and R. Solé. The role of computation in complex regulatory networks. In E. Koonin et al, editor, Scale-free Networks and Genome Biology, 206–225. Landes Bioscience, 2006.

[43] R. Ferrer-i-Cancho and R. V. Solé. Least effort and the origins of scaling in human language. Proc. Natl. Acad. Sci. USA, 100:788–791, 2003.

[44] W. Fontana. Algorithmic Chemistry. In C. G. Langton, C. Taylor, J. D. Farmer, and S. Rasmussen, editors, Artificial Life II, 159–210, Redwood City, CA, 1992. Addison-Wesley.

[45] J. Fortuny and B. Corominas-Murtra. Some formal considerations on the generation of hierarchically structured expressions. Catalan Journal of Linguistics, 8:99–111, 2009.

[46] X. Gabaix. Zipf's law for cities: An explanation. The Quarterly Journal of Economics, 114(3):739–767, 1999.

[47] M. Gell-Mann and S. Lloyd. Information measures, effective complexity, and total information. Complexity, 2(1):44–52, 1996.

[48] H. Goldstein, C. P. Poole, and J. L. Safko. Classical Mechanics (3rd Edition). Addison Wesley, 3 edition, 2001.

[49] P. D. Grünwald and P. M.B. Vitányi. Kolmogorov complexity and information theory with an interpretation in terms of questions and answers. J. of Logic Language and Information, 12:497–529, 2003.

[50] P. D. Grünwald and P. M.B. Vitányi. Handbook of the Philosophy of Science, Volume 8: Philosophy of Information., chapter: Algorithmic Information Theory, pages 289–325. Elsevier Science Publishers, 2008.

[51] H. Haken. Synergetics: An Introduction. Nonequilibrium Phase Transitions and Self- Organization in Physics, Chemistry and Biology (Springer Series in Synergetics). Springer, 1978.

[52] P. Harremoës and F. Topsøe. Maximum entropy fundamentals. Entropy, 3:191–226, 2001.

[53] P. Harremoes and F. Topsoe. Zipf's law, hyperbolic distributions and entropy loss. In Proceedings of the IEEE International Symposium on Information Theory, page 207, Lausanne, 2002. IEEE.

[54] M. D. Hauser, N. Chomsky, and T. W. Fitch. The faculty of language: What is it, who has it, and how did it evolve? Science, 298:1569–1579, 11 2002.

[55] C. Hockett. Language, mathematics and linguistics. Current Trends in Linguistics: Theoretical Foundations, 3:155–304, 1966.

[56] F. Homae, H. Watanabe, T. Nakano, and G. Taga. Prosodic processing in the developing brain. Neurosci Res, 59(1):29–39, 2007.

[57] J. Hopcroft and J. Ullman. Introduction to Automata Theory, Languages and Computation. Addison-Wesley. New York, 1979.

[58] J. Hopfield. Physics, computation, and why biology looks so different. Journal of Theoretical Biology, 171(1):53–60, 1994.

[59] W. Horsthemke and R. Lefever. Noise-Induced Transitions: Theory and Applications in Physics, Chemistry, and Biology (Springer Series in Synergetics). Springer, 1983.

[60] Z.-F. Huang and S. Solomon. Power, levy, exponential and gaussian regimes in autocatalytic financial systems, 2000.

[61] J. Hurford. Biological evolution of the saussurean sign as a component of the language acquisition device. Lingua, 77(2):187–222, 1989.

[62] E. T. Jaynes. Information theory and statistical mechanics. Physical Review, 106(4):620–630, 1957.

[63] D. S. Jones. Elementary Information Theory. Oxford University Press. Oxford, 1979.

[64] J. N. Kapur. Maximum-entropy Models in Science and Engineering. New Delhi. Wiley Eastern Limited, 1989.

[65] S. A. Kauffman. The Origins of Order: Self-Organization and Selection in Evolution. Oxford University Press, 1993.

[66] J. Kelley. General Topology. Van Nostrand, 1955.

[67] A. I. Khinchin. Mathematical Foundations of Information Theory. Dover Publications, 1957.

[68] A. Kolmogorov. Logical basis for information theory and probability theory. Problems Inform. Transmission, 1:1–7, 1965.

[69] N. L. Komarova and P. Niyogi. Optimizing the mutual intelligibility of linguistic agents in a shared world. Art. Int., 154(1-2):1–42, 2004.

[70] S. Kripke. Wittgenstein on Rules and Private Language. Basil Blackwell Publishing, 1982.

[71] P. Krugman. Confronting the mystery of urban hierarchy. J. Jap. Int. Econ., 10:399–418, 1996.

[72] S. Kullback. Information Theory and Statistics. John Wiley and Sons. New York, 1959.

[73] G. Lebon, D. Jou, and J. Casas-Vázquez. Understanding Non-equilibrium Thermodynamics: Foundations, Applications, Frontiers. Springer-Verlag:Berlin, 2008.

[74] W. Li. Random texts exhibit zipf's-law-like word frequency distribution. Information Theory, IEEE Transactions on, 38(6):1842–1845, 1992.

[75] W. R. Lowenstein. The touchstone of life. Molecular information, cell communication and the foundations of life. Oxford University Press. New York., 1993.

[76] H. A. Makse, S. Havlin, and H. E. Stanley. Modelling urban growth patterns. Nature, 377(6550):608–612, 1995.

[77] O. Malcai, O. Biham, and S. Solomon. Power-law distributions and lévy-stable intermittent fluctuations in stochastic systems of many autocatalytic elements. Phys. Rev. E, 60(2):1299–1303, 1999.

[78] S. C. Manrubia and D. H. Zanette. Intermittency model for urban development. Phys. Rev. E, 58(1):295–302, 1998.

[79] R. Margalef. Perspectives in ecological theory. University of Chicago Press, Chicago, 1968.

[80] J. Maynard-Smith and E. Szathmàry. The Major Transitions in Evolution. University of New York Press. New York, 1997.

[81] G. A. Miller and N. Chomsky. Finitary models of language users. In R. D. Luce, R. Bush, and E. Galanter, editors, Handbook of Mathematical Psychology, volume 2, 419–491. Wiley, New York, 1963.

[82] M. Li and P.l Vitányi. An introduction to Kolmogorov complexity and its applications. Springer, New York [u.a.], 1997.

[83] J. Von Neumann. Probabilistic logics and the synthesis of reliable organisms from unreliable components. In C. E. Shannon and J. McCarthy, editors, Automata Studies. Princeton University Press. Princeton, 1956.

[84] M. E. J. Newman. Power laws, pareto distributions and zipf's law. Contemporary Physics, 46, 2005.

[85] M. E. J. Newman and Kim Sneppen. Avalanches, scaling, and coherent noise. Phys. Rev. E, 54(6):6226–6231, Dec 1996.

[86] E. L. Newport. Maturational constraints on language learning. Cogn. Sci., 14(1):11–28, 1990.

[87] M. A. Nowak and D. Krakauer. The evolution of language. Proc. Nat. Acad. Sci. USA, 96(14):8028–8033, 1999.

[88] V. Pareto. Cours de d'economie Politique. Droz. Geneva, 1896.

[89] R. K. Pathria. Statistical Mechanics, Second Edition. Butterworth-Heinemann, 1996.

[90] A. R. Plastino, H. G. Miller, and A. Plastino. Minimum kullback entropy approach to the fokker-planck equation. Phys. Rev. E, 56(4):3927–3934, 1997.

[91] A. Radford. Syntactic Theory and the Acquisition of English Syntax: the nature of early child grammars of English. Oxford. Blackwell, 1990.

[92] A. Radford. Syntax: A minimalist introduction. Cambridge University Press. Cambridge, 1997.

[93] J. Rosselló. Combinatorial properties at the roots of language. In Joana Rosselló and Txuss Martín, editors, The Biolinguistic Turn. Issues on Language and Biology, 162–186. Promociones y Publicaciones Universitarias, S.A., 2006.

[94] F. Saussure. Cours de Linguistique Générale. Bibliothèque scientifique Payot: Paris, 1916.

[95] C. E. Shannon. A mathematical theory of communication i. Bell Sysytem Technical Journal, 27:379–423, 1948.

[96] H. A. Simon. On a class of skew distribution functions. Biometrika, 42:425–440, 1955.

[97] R. Solé, B. Corominas-Murtra, and J. Fortuny. Diversity, competition, extinction: the ecophysics of language change. Journal of The Royal Society Interface, 2010.

[98] R. Solé, B. Corominas-Murtra, S. Valverde, and L. Steels. Language networks: Their structure, function, and evolution. Complexity, 15(6):20–26, 2010.

[99] S. Solomon and M. Levy. Spontaneous scaling emergence in generic stochastic systems. Int. J. Mod. Phys. C, 7:745–751, 1996.

[100] R. Solomonoff. A formal theory of inductive inference. Inform. and Control, 7-1:1–22, 1964.

[101] P. Suppes. Axiomatic Set Theory. Dover: New York, 1972.

[102] P. Vogt. Minimum cost and the emergence of the zipf-mandelbrot law. In In J. Pollack, M. Bedau, P. Husbands,

T. Ikegami, and R. A. Watson, editors, Artificial Life IX Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems. MIT Press, 2004.

[103] W. H. White. On the form of steady-state solutions to the coagulation equations. J. Colloid. Interface Sci., 87:204–208, 1982.

[104] N. Wiener. Cybernetics. John Wiley, New York, 1948.

[105] L. Wittgenstein. Philosophical Investigations. Blackwell Publishing, 1953/2001.

[106] H. P. Yockey. Information Thery and Molecular Biology. Cambridge University Press, Cambridge (UK), 1992.

[107] D. H. Zanette and S. C. Manrubia. Role of intermittency in urban development: A model of large-scale city formation. Phys. Rev. Lett., 79(3):523–526, 1997.

[108] G. K. Zipf. Human Behavior and the Principle of Least Effort. Addison-Wesley (Reading MA), 1949.

Chapter 5

PAPERS ON LANGUAGE ACQUISITION

Solé RV, Corominas-Murtra B, Valverde S, Steels L. Language networks: their structure, function and evolution. Complexity. 2010; 15(6): 20-6.

Advances in Complex Systems

THE ONTOGENY OF SCALE-FREE SYNTAX NETWORKS: EVIDENCE FOR CHOMSKY'S HIDDEN HARDWARE?

BERNAT COROMINAS-MURTRA
ICREA-Complex Systems Lab, Universitat Pompeu Fabra (GRIB), Dr Aiguader 80, 08003 Barcelona, Spain
bernat.corominas@upf.edu

SERGI VALVERDE
ICREA-Complex Systems Lab, Universitat Pompeu Fabra (GRIB), Dr Aiguader 80, 08003 Barcelona, Spain
sergi.valverde@upf.edu

RICARD SOLÉ
ICREA-Complex Systems Lab, Universitat Pompeu Fabra (GRIB), Dr Aiguader 80, 08003 Barcelona, Spain
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA
ricard.sole@upf.edu

Received (received date)
Revised (revised date)

Language development in children provides a window to understanding the transition from protolanguage to language. Here we present the first analysis of the emergence of syntax in terms of complex networks. A previously unreported, sharp transition is shown to occur around two years of age from a (pre-syntactic) tree-like structure to a scale-free, small world syntax network. The nature of such transition supports the presence of an innate component pervading the emergence of full syntax. This observation is difficult to interpret in terms of any simple model of network growth, thus suggesting that some internal, perhaps innate component was at work.

Keywords: Language evolution, language acquisition, syntax, complex networks, small worlds

1. Introduction

The process of acquiring language during childhood is a remarkable one. Children start learning and babbling single words but at some point two-word combinations are made. Vocabulary size rapidly increases and at some point the child starts mastering the grammar . Grammatical rules define the fabric of language and they include phonology (how words are formed by putting sounds together), syntax (how words combine to form phrases) and semantics (which allows interpreting the meaning of words and sentences). The process is nonlinear: although the size of the lexicon

grows in a monotonous fashion, the organization of all components of grammar does not change smoothly [38]. At a certain age, rapid shifts are observed: children start using complex sentences with well-organized grammatical structure. The previous stage is two-word sentences and no “three-word” phase seems to exist. What is the meaning of these sudden changes?

Language acquisition is not only a problem of language development: it actually provides a window into language origins and its evolution. The origins of complex forms of communication [28] are hotly debated [8, 45]. One obvious problem of this is the lack of fossils [2] which forces us to use alternative, indirect sources of information. An important issue here is the mechanisms involved in shaping language. Natural selection appears as an essential candidate [22, 34, 26]. On the other hand, some authors suggest that language is a byproduct of a large brain, with neural structures formerly used for other functions [19]. A different view suggests that language must be considered a complex “organism” in itself [7]. Such organism would change on faster time scales than those related to genetic change. Moreover, computational models using robotic agents can also help understanding potential scenarios for the emergence of communication in artificial communities [43] not necessarily tied to selective forces.

Confronted with the surprising mastery of complex grammar achieved by children over two years, some authors early concluded that an innate, hardwired element (a “language acquisition device”) must be at work [6, 37, 36]. Children are able to construct complex sentences by properly using phonological, syntactic and semantic rules in spite that no one teaches them[a]. Steven Pinker has usedthe term “language instict” comparing the capacity of using grammar with how spiders “know” how to spin webs [36]:

> Web-spinning was not invented by some unsung spider genius and does not depend on having had the right education or on having an aptitude for architecture or the construction trades. Rather, spiders spin spider webs because they have spider brains, which give them the urge to spin and the competence to succeed.

The metaphor is useful but perhaps too strong. Spiders do not need any training from adults in order to create their designs. Instead, children must receive input from the social environment in order to gather the necessary components that are required to trigger (to the least) the emergence of a complex language. But how can children acquire such huge set of rules? Are there some specific, basic rules predefined as a part of the biological endowment of humans? If so, some mechanism of language acquisition (the universal grammar, UG) should guide the process. In this way, models assuming a constrained set of accessible grammars have shown that final states (i.e., an evolutionary stable complex grammar) can be reached under a

[a]Specifically, they can generate a virtually infinite set of grammatically correct sentences in spite that they have been exposed to a rather limited number of input examples

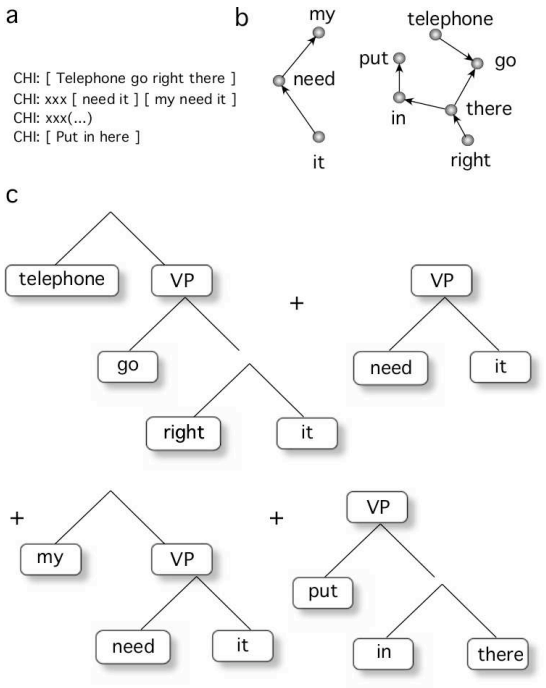


Fig. 1. Building the networks of Syntax Acquisition. First we identify the structures in child's productions (a) using the lexico-thematic nature of early grammars [38], see [9]. Afterwards, a basic constituency analysis is performed (b) assuming that the semantically most relevant item is the head of the phrase and that the verb in finite form (if any) is the head of the sentence. Finally (c) a projection of the constituent structure in a dependency graph is obtained.

limited exposure to the right inputs [25, 33]. However, we cannot deny the fact that important features of the language acquisition process can be obtained by appealing only to general purpose mechanisms of learning [32, 12, 27] or the importance of pure self-organization in the structure of the speech code [42, 35]. An integrated picture should take into account the interaction of some predefined grammar features with general purpose mechanisms of learning and code self-organization, structuring human languages as we know today. Under this view, transition from protogrammar to grammar would be the result of an innovation of brain organization rapidly predated for communication [19].

A quantitative analysis of language acquisition data is a necessary source of validation of different hypotheses about language origins and organization. Indeed, it is well accepted that any reasonable theory of language should be able to explain

how it is acquired. Here we analyze this problem by using a novel approximation to language acquisition based on a global, network picture of syntax. Instead of following the changes associated to lexicon size or counting the enumber of pairs (or strings) of words, we rather focus on how words relate to each other and how this defines a global graph of syntactic links. We focus our analysis in the presence of marked transitions in the global organization of such graphs. As shown below, both the tempo and mode of network change seem consistent with the presence of some predefined hardware that is triggered at some point of child's cognitive development. Furthermore, we explore this conjecture by means of an explicit model of language network change that is able to capture many (but not all) features of syntax graphs. The agreements and disagreements can be interpreted in terms of non-adaptive and adaptive ingredients of language organization.

2. Building syntactic Networks

Language acquisition involves several well-known stages [38]. The first stage is the so-called *babbling*, where only single phonemes or short combinations of them are present. This stage is followed by the *Lexical spurt*, a sudden lexical explosion where the child begins to produce a large amount of isolated words. Such stage is rapidly replaced by the *two words stage*, where short sentences of two words are produced. In this period, we do not observe the presence of functional items nor inflectional morphology. Later, close to the two-years age, we can observe the *syntactic spurt*, where more-than-two word sentences are produced. The data set studied here includes a time window including all the early, key changes in language acquisition, from non-grammatical to grammatical stages.

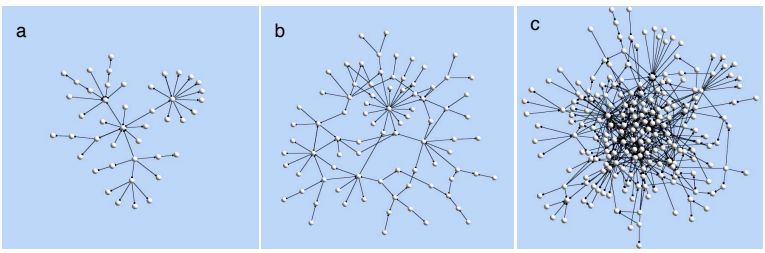


Fig. 2. Qualitative view of the transitions from tree-like graphs to scale-free syntax graphs through the acquisition process. Here three snapshots of the process are shown, at (a) about 22 months, (b) about 23 moths and (c) about 25 months. Although a tree-like structure is shown to be present through the pre-transition (a-b) a scale-free, much more connected web suddenly appears afterward (c), just a month later.

In this paper we analyse raw data obtained from child's utterances, from which we extract a global map of the pattern of the use syntactic relations among words.

In using this view, we look for the dynamics of large-scale organization of the use of syntax. This can be achieved by means of complex networks techniques, by aggregating all syntactic relationships within a graph. Recent studies have shown that networks reveal many interesting features of language organization [29, 16, 39, 14, 21, 24] at different levels. These studies uncovered new regularities in language organization but so far none of them analyzed the emergence of syntax through language acquisition. Here we study in detail a set of quantitative, experimental data involving child utterances at different times of their development.

A very interesting contribution to this problem is provided by the recent work of Jinyun Ke and Yao Yao on language development using co-occurrence graphs (Ke and Yao, 2008). These authors used conversations from recorded sessions where a child speaks with adults spontaneously. They used the Manchester corpus from *CHILDES* database [b]. In their study, Ke and Yao considered pairs of words as linked provided that they appear collocated within at least one utterance. By looking at the whole set of relations between words, a network of language acquisition was generated. It was found that children paths of aquisition where different from individual to individual. However, a remarkable regularity was also found: children with smaller vocabularies displayed a higher flexibility in word combination, indicating that the global organization of word interactions plays a nontrivial role. Children used in this study where typically older than two years and thus some key phenomena around this critical age could not be studied. This is the target of our analysis.

In this work we consider a different class of network the so called the *syntax network* $\mathcal{G} = \mathcal{G}(\mathcal{W}, E)$ defined as follows (see fig.1). Using the lexicon at any given acquisition stage, we obtain the collection of words $W_i (i = 1, ..., N_w)$, being every word a node $w_i \in \mathcal{G}$. There is a connection between two given words provided that they are syntactically linked[c]. The set of links E describes all the syntactic relationships in the corpus. For every acquisition stage, we obtain a syntactic network involving all the words and their syntactic relationships. The structure of syntax networks will be described by means of the *adjacency matrix* $A = [a_{ij}]$ with $a_{ij} = 1$ when there is a link between words w_i and w_j and $a_{ij} = 0$ otherwise.

Specifically, we choose Peter's corpora [3, 4] and Carl's corpora, from the Manchester corpus [47] as particularly representative and complete examples. Time intervals have been choosen to be regular and both sets of corpora span a time window that can be considered large enough to capture statistically relevant properties. Each corpus contains several conversations among adult investigators and the corresponding child. However, the raw corpus must be parsed in order to construct properly defined graphs. In [9] we present a detailed description of the criteria and rules followed to pre-process the raw data. The main features of the parsing algo-

[b] http://talkbank.org

[c] Recall that the net is defined as the projection of the constituency hierarchy. Thus, the *link* has not an ontological status under our view of syntax[9]

rithm are indicated in fig.1 and can be summarized as follows:

(1) Select only child's productions rejecting imitations, onomatopoeia's and undefined lexical items.
(2) Identify the *structures*, i.e., the minimal syntactic constructs.
(3) Among the selected structures, we perform a basic analysis of constituent structure, identifying the verb in finite form (if any) in different phrases.
(4) Project the constituent structures into lexical dependencies. This projection is close to the one proposed by [21] within the framework of the network-based *Word Grammar*[d].
(5) Finally, we build the graph by following the dependency relations in the projection of the syntactic structures found above. Dependency relations allow us to construct a syntax graph.

With this procedure, we will obtain a graph for every corpus. The resulting graphs will be our object of study in the following section.

3. Evolving syntax Networks

Here we analyze the topological patterns displayed by syntax networks at different stages of language acquisition. To our knowledge, this is the first detailed analysis of language network ontogeny so far. The resulting sequence exhibits several remarkable traits. In fig. (2) we show three examples of these networks. At early stages, (fig. 2a,b) most words are isolated (not shown here) indicating a dominant lack of word-word linkage. Isolated words are not shown in these plots. For each stage, we study the largest subset of connected words or *giant component* (GC). The reason for considering the largest connected component is that, from the very beginning, the GC is much larger than any other secondary connected component and in fact the system shows an almost all-or-none separation between isolated words and those belonging to the GC. In other words, the giant component captures almost all word-word relations. By sampling a set of corpora at different times, we obtain a time series of connected networks $\mathcal{G}(\mathcal{W}_T, E_T)$, where $\mathcal{W}_T$ and E_T are the set of words and links derived from the T-th corpus.

3.1. *Global organization*

In agreement with the well-known presence of two differentiated regimes, we found that networks before the two-year transition (fig.2a-b) show a tree-like organization, suddenly replaced by much larger, heterogeneous networks (fig.2c) which are very similar to adult syntactic networks [14]. The gray area indicates the presence of complex syntactic organization (massive presence of structures with more than two words). This abrupt change indicates a global reorganization marked by a shift in

[d]note that the operation is reversible, since can rebuild the tree from the dependency relations

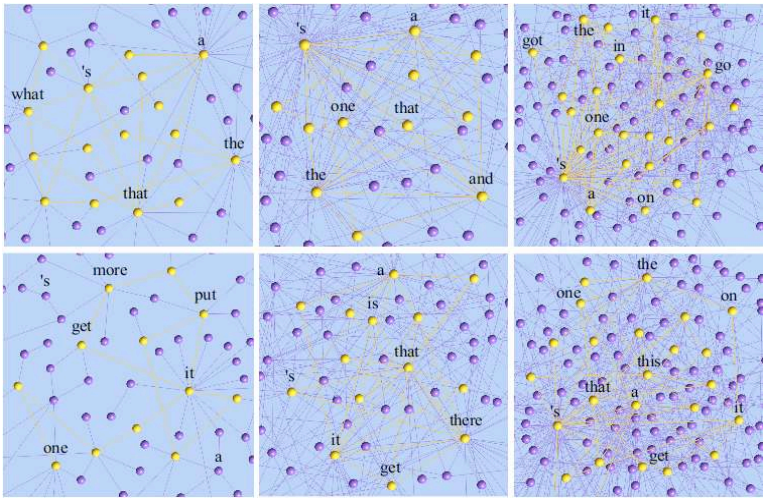


Fig. 3. Evolution of the organization of hubs and their nature. There is a critical change at the two-years age marked by a strong reorganization of the network. In Carl's set of corpora (above), we observe a strong reinforcement of their connectiovity. However, in Peter's set of corpora (down) we observe, prior to the transition, that semantically degenerated elements (such as *it*) act as hubs. Key words essential to adult syntax are missing in these early stages. After the transition, the hubs change from semantically degenerated to functional items (i.e., *a* or *the*). The ages of these snapshots correspond, for Carl, to 1 year and 9 months, 1 year and 11 months and 2 years and 2 months, respectivley. For Peter, the ages of the snapshots correspond to 1 year and 11 months, 2 years and 2 years and 3 months.

grammar structure. Both Peter's and Carl's corpora exhibit these two clearly differentiated regions. Furthermore, in Peter's set of corpora, we can observe explicitly, another qualitative change. Indeed, when looking to the changes in the nature of hubs before and after the transition we see that highly connected words in the pre-transition stage are semantically degenerated lexical items, such as *it*. After the transition, hubs emerge as functional items, such as *a* or *the*. Carl's corpora exhibit the presence of the functional particles as hubs from the very begining. However, the hubs of the pre-transition stage are notably weaker than after the transition.

3.2. *Small world development*

Two important measures allow us to characterize the overall structure of these graphs. These are the average path length L_T and clustering coefficient C_T [50]. The first measure is defined as the average $D_T = \langle D_{min}(i,j)\rangle$, where $D_{min}(i,j)$ indicates the length of the shortest path connecting nodes w_i and w_j. The average is performed over all pairs of words. Roughly speaking, short path lengths means

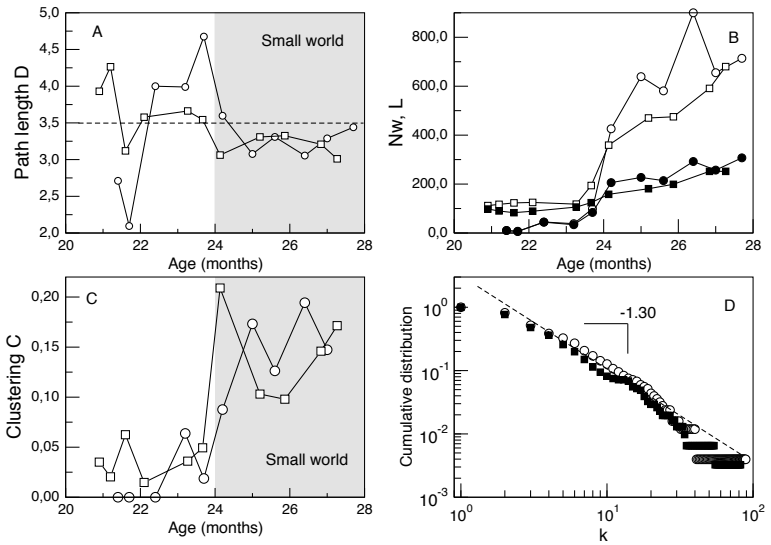


Fig. 4. Changes in the structure of syntax networks in children are obtained by means of several quantitative measures associated to the presence of small world and scale-free behavior. Here we display: (a) the average path length D_T, (b) The number of words (N_w) and links L (c) the clustering coefficient. As shown in (a) and (c), a small world pattern suddenly emerges after an age of ≈ 24 months. A rapid transition from a large L and low C takes place towards a small world network (with low D and high C). After the transition, well-defined scale-free graphs, with $P(k) \propto k^{-2.30}$, are observed (d).

that it is easy to reach any given word w_i starting from another arbitrary word w_j. Small path lengths in sparse networks are often an indication of efficient information exchange. The clustering coefficient C_T is defined as the probability that two words that are neighbors of a given word are also neighbors of each other (i. e. that a triangle is formed). In order to estimate C_T, we define for each word w_i a neighborhood Γ_i. Each word $w_j \in \Gamma_i$ is syntactically related (at least once) with w_i in a production. The words in Γ_i can also be linked to each other, and the clustering $C(\Gamma_i)$ is defined as

$$C(\Gamma_i) = \frac{1}{k_i(k_i-1)} \sum_j \sum_{k \in \Gamma_i} a_{jk} \tag{1}$$

The average clustering of the G_T network is simply $C_T = \langle C(\Gamma_i) \rangle$ i.e, the average over all $w_i \in W$. Most complex networks in nature and technology are known to be *small words*, meaning that they have short path lengths and high clustering [50] Although language networks have been shown to have small world structure [16, 44, 14, 39] little is known about how it emerges in developing systems.

Two regimes in language acquisition can be also observed in the evolution of the average path length fig.(4a). It grows until reaches a peak at the transition (where the small word domain is indicated by means of the grey area). Interestingly, about $T = 5$ both networks display the highest number of words for the pre-transition stage. For $T > 5$, the average path length stabilizes to $D_T \approx 3.5$ for Peter's set of corpora and $D_T \approx 3.1$ in Carl's one (see fig. (4 b)). The increasing trend of D_T in $T < 5$ may be an indication that combinatorial rules are not able to manage the increasing complexity of the lexicon. In fig.(4b) we plot the corresponding number of words N_T and links L_T of the GC as filled and open circles, respectively. We can see that the number of connected words that belong to the GC increases in a monotonous fashion, displaying a weak jump at the age of two. However, the number of links (and thus the richness of syntactic relations) experiences a sharp change.

The rapid increase in the number of links indicates a qualitative change in network properties strongly tied to the reduction of the average path length. A similar abrupt transition is observed for the clustering coefficient: In the pre-transition stage C_T are small (zero for $T = 1, 2, 3$, in Peter's set of corpora). After the transition, both sets of corpora exhibit a sudden jump to converge around $C_T \approx 0.16$. Both D_T and C_T are very similar to the measured values obtained from syntactic graphs from written corpus [14].

3.3. *Scale-free topology*

The small world behavior observed at the second phase is a consequence of the heterogeneous distribution of links in the syntax graph. Specifically, we measure the degree distribution $P(k)$, defined as the probability that a node has k links. Our syntactic networks display scale-free degree distributions $P(k) \propto k^{-\gamma}$, with $\gamma \approx 2.3 - 2.5$. Scale-free webs are characterized by the presence of a few elements (the hubs) having a very large number of connections. Such heterogeneity is often the outcome of multiplicative processes favouring already degree-rich elements to gain further links [1, 10, 11].

An example is shown in fig.(4d) where the cumulative degree distribution, i.e:

$$P_>(k) = \int_k^\infty P(k)dk \sim k^{-\gamma+1} \tag{2}$$

is shown. In both collection of nets, the fitting gives a very close scaling exponent $\gamma \approx 2.3$, also in agreement with adult studied corpora. They are responsible for the very short path lengths and thus for the efficient information transfer in complex networks. Moreover, relationships between hubs are also interesting: the syntax graph is *dissassortative* [31], meaning that hubs tend to avoid to be connected among them [14]. In our networks, this tendency also experiences a sharp change close to the transition domain (not shown) thus indicating that strong constraints emerge strongly limiting the syntactic linking between functional words.

4. Null Models of Network growth

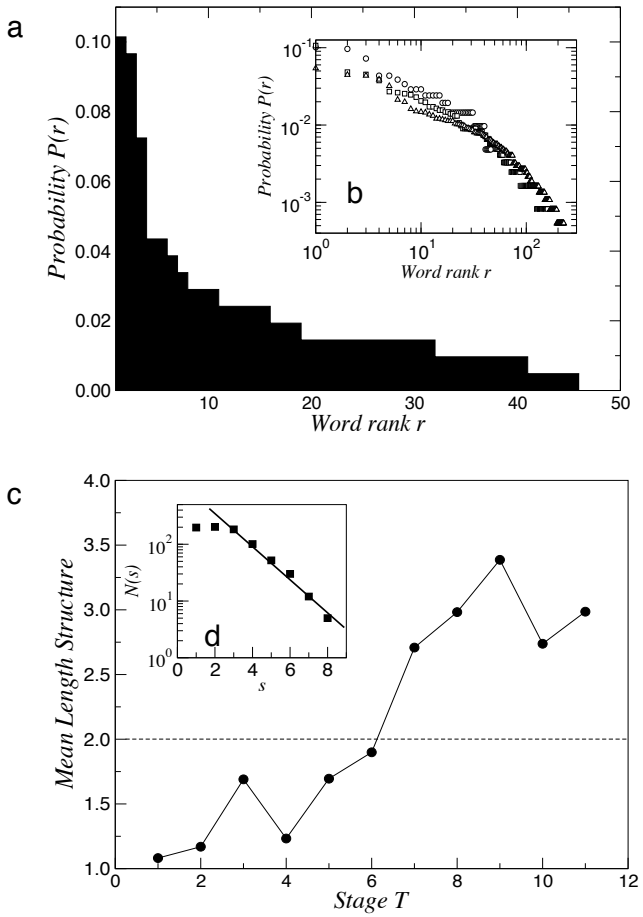


Fig. 5. Statistical patterns in language acquisition. In (a) an example of the rank-frequency distribution of lexical items is shown (here for Peter's corpus (see text) at stage $T = 2$ (1 year and 10 months)). The inset (b) displays three examples of such skewed distributions in log-log scale for $T = 2$ (circles), $T = 5$ (squares) and $T = 8$ (triangles). In (c) the evolution of mean length of structure (L) is displayed. It gives an estimate of the (linear) complexity of the productions generated at different stages. The dashed line indicates the two word production size. After stage $T = 5$, the MSL ($\langle s \rangle$, in the text) comes close to two and a sharp change occurs. In (d) we also show an example of the frequency distribution $N(L)$ for these productions in linear-log form for $T = 5$.

We have described a pattern of change in syntax networks. The patterns are nontrivial and quantitative. What is their origin? Can we explain them in terms of some class of self-organization (SO) model? Are they instead associated to some internal, hardwired component? Here we present a new model of network evolution that tries to capture the observed changes and provides tentative answers to the previous questions.

4.1. *Simple SO graph growth models*

We explored several types of SO models without success. Appropriate models should be able to generate: (a) sharp changes in network connectivity and (b) scale-free graphs as the final outcome of the process. In relation to the sudden shift, it is well known that a sharp change in graph connectivity occurs when we add links at random between pairs of nodes until a critical ratio of links against nodes is reached [13, 5]. Starting from a set of N isolated elements, once the number of links L is such that $p \equiv L/N \approx 1$, we observe a qualitative change in graph structure, from a set of small, separated graphs ($p < 1$) to a graph structure displaying a giant component ($p > 1$) with a comparatively small number of isolated subgraphs. This type of *percolation* model has been widely used within the context of SO [23, 40]. Unfortunately, such a transition is not satisfactory to explain our data, since (a) it gives graph with a Poissonian degree distribution [5], i.e.

$$P(k) \approx \frac{\langle k \rangle^k e^{-k}}{k!} \tag{3}$$

and (b) there is no sharp separation between isolated nodes and a single connected graph, but instead many subgraphs of different sizes are observed.

Other models instead consider growing graphs using preferential attachment rules [1, 10, 11]. In these models the number of nodes grows by adding new ones which tends to link with those having the largest connectivity (a rich-gets-richer mechanism). Under a broad range of conditions these amplification mechanisms generate scale-free graphs. However, the multiplicative process does not lead to any particular type of transition phenomenon. The status of hubs remains the same (they just win additional links). Actually, well-defined predictions can be made, indicating that the degree of the hubs scales with time in a power-law form [1, 10].

Although many possible combinations of the previous model approaches can be considered, we have found that the simultaneous presence of both scale-free structure emerging on top of a tree and a phase transition between both is not possible. In order to properly represent the dynamics of our network, a data-driven approach seems necessary.

4.2. *Network growth model and analysis*

In order to reproduce the observed trends, we have developed a new model of network evolution. The idea is to describe the process of network growth without

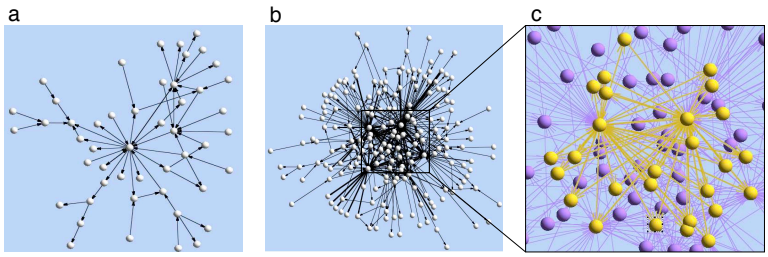


Fig. 6. Sudden changes in network organization from the language acquisition model (see text). In (a) and (b) we display the largest subgraph before (c) and right after (b) the transition. The graphs share the basic change from tree-like to scale-free structure, although exhibit higher clustering coefficients. In (c) a blow-up of (b) is shown, indicating the presence of a few hubs that are connected among them both directly and through secondary connectors.

predefined syntactic rules. We make the simplistic assumption that word interaction only depends on word frequency following Zipf's law. In this context, it has been suggested that Zipf's law might be the optimal distribution compatible with efficient communication [18, 17, 15, 41]. If no internal mechanisms are at work, then our model should be able to capture most traits of the evolution of syntax.

For the sake of simplicity, and due to the similarity of the two sets of data, we base our comparative study with the Peter's set of corpora. In order to develop the model, a new measure, close to the usual MLU[e]used in linguistics, must be defined. The *structure length* of the i-th structured production (s_i) is measured by counting the number of words that participate in the i-th syntactic structure. In our previous example (see figure 1) we had 4 structures, of sizes $|s_1| = 4, |s_2| = 2, |s_3| = 2$ and $|s_4| = 3$. Its average, the *Mean Structure Length,* $\langle s \rangle$ is $\langle s \rangle = 2.75$. In fig. (5-c) we can see how the MSL evolves over time. The frequency of s, $p(s)$ was also measured and was found to decay exponentially, with $p(s) \propto e^{-|s|/\gamma}$, with $\gamma = 1.40$ in this specific set of data (fig. (5-d)). We can connect the two previous through

$$\langle s \rangle = \frac{1}{Q} \sum_s s e^{-|s|/\gamma} \tag{4}$$

where Q is defined as the normalization constant:

$$Q = \sum_s e^{-|s|/\gamma} \tag{5}$$

In the five first corpora, $\langle s \rangle < 2$. Beyond this stage, it rapidly grows with $\langle s \rangle > 2$, (see fig. (5-b)).

[e]The MLU is the *Mean Length of Utterance* i.e. the average length of a child's utterances, measured in either words or morphemes.

We incorpore to the data-driven model our knowledge on structure lengths. We first construct, for each corpus, a random syntactic network that shares the statistics of word frequencies and structure lengths of the corresponding data set. Such a measure can be interpreted, in cognitive terms, as some kind of working memory and might be the footprint of some maturational constraints [32, 12].

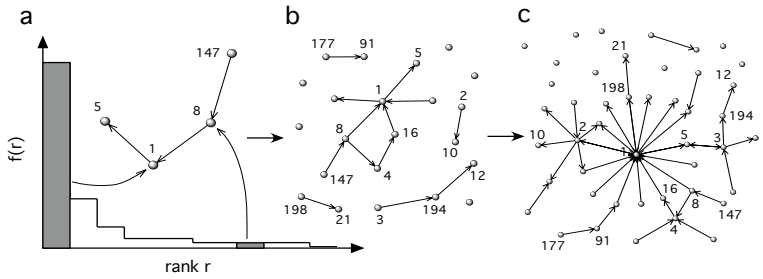


Fig. 7. Algorithm for network growth. The model uses as input information a Zipf's distribution of “words” and the probability to find a structure of size s in a given corpus, $p_T(s)$. Each step we choose s words from the list, each word with a probability proportional to their frequency. A link is then established between two successive words generating an unstructured string of s nodes. We repeat the process a number of times and we aggregate in a global graph all the obtained strings. $p_T(s)$ can be interpreted as the footprint of a kind of working memory, and follows an exponential distribution (As shown in fig. (5))

For simplicity, we assume that the probability of the i-th most frequent word is a scaling law:

$$p_w(i) = \frac{1}{Z} i^{-\beta} \tag{6}$$

with $1 \leq i \leq N_w(T)$, $\beta \approx 1$ and Z is the normalization constant:

$$Z = \sum_{i=1}^{N_w(T)} \left(\frac{1}{i}\right)^{\beta} \tag{7}$$

(notice that Z depends on lexicon size, $N_w(T)$, which grows slowly at this stage). However, the actual word frequency is affected by other corpus features. In particular, our corpora are highly redundant with many duplicated structures but we build our nets ignoring such redundancies, since we are interested in the topological patterns of use. For every corpus T with $N_s(T)$ distinct structures, we compute the distribution of structure lengths $p_T(s)$, $1 \leq T \leq 11$. From $N_w(T)$, $p_w(i)$, $N_s(T)$ and $p_T(s)$, we generate a random syntactic network for every stage $1 \leq T \leq 11$ (see fig.(7)). Given a lexicon with $N_w(T)$ different items, labeled as $a_1 ... a_{N_w(T)}$ the model algorithm goes as follows:

(1) Generate a random positive integer s with probability $p_T(s)$.

(2) Choose s different "words" from the lexicon, $a_k^1, ..., a_j^s$ each word with probability $p(a_i) \propto i^{-\beta}$, with $\beta \approx 1$.
(3) Trace an arc between every two successive words thus generating a unstructured string of s nodes.
(4) Repeat (1), (2) and (3) until $N_s(T)$ structures are generated.
(5) Aggregate all the obtained strings in a single, global graph.

In spite of the small number of assumptions made, the above model reproduces many of the topological traits observed in real networks. To begin with, we clearly observe the sudden transition from tree-like networks to scale-free networks (see fig.6). Furthermore, typical network properties, such as clustering, degree distribution or path lenghts seem to fit real data successfully (see fig. (8)). The very good agreement between global patterns of network topology is remarkable given the lack of true syntax. It indicates that some essential properties of syntax networks come "for free". In other words, both the small world and the scale-free architecture of syntax graphs would be spandrels: although these type of networks provide important advantages (such as highly efficient and robust network interactions) they would be a byproduct of Zipf's law and increased neural complexity. These results thus support the non-adaptive nature of language evolution.

However, particularly beyond the transition, a detailed analysis is able to find important deviations between data and model predictions. This becomes specially clear by looking at small subgraphs of connected words. Studying small size subgraphs allows to explore local correlations among units. Such correlations are likely to be closer to the underlying rules of network construction, since they are limited specificaly to direct node-node relations and their frequency. We have found that the subgraph census reveals strong deviations from the model due to the presence of grammatical constraints, i.e, non-trivial rules to build the strings.

In figure (9) we display the so-called subgraph census plot [20, 49] for both real (circles) and simulated (squares) networks. Here the frequencies of observed subgraphs of size three are shown ordered in decreasing order for the real case. For the simulated networks, we have averaged the subgraph frequencies over 50 replicas. Several obvious differences are observed between both censuses. The deviations are mainly due to the hierarchical relations that display a typical syntactic structure, and to the fact that lexical items tend to play the same specific role in different structures (see fig.9b-d). Specifically, we find that the asymetries in syntactic relations induce the overabundance of certain subgraphs and constrain the presence of others. Specially relevant is the low value of third type of subgraph, confronted with the model prediction. This deviation can be due to the *organizing* role of functional words (mainly out-degree hubs) in grammar. Indeed, coherently with this interpretation, we find that the first type of subgraph (related with out-degree hubs) is more abundant than the model prediction.

The second interesting deviation within this set of corpora, is given by the changing status of hubs. As previously described, in the prefunctional period hubs

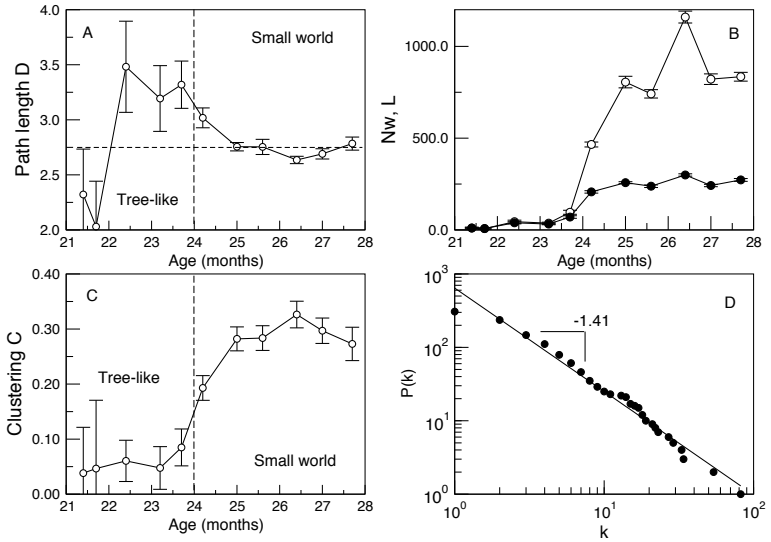


Fig. 8. Changes in the structure of syntax model networks -compare with fig.(4). Here we show: (a) the average path length D, (b) the number of links (DL) and lexical items (N) and (c) the clustering coefficient C. An example of the resulting SF distributions is also shown in (d).

are semantically degenerated words, such as *that, it*, whereas beyond the transition hubs are functional words. This observation seems to be coherent with a recently proposal to understand the emergence of functional items in child grammars. In short, a pure articulattory strategy introduces a new sound (mainly the *a*) that is rapidly predated by the syntactic system when it is mature enough [48]. This would imply a reuse of an existing, phonetical element and would explain the astonishing increasing of appearance that they experience. If we follow the changes in number of links displayed by the hubs in the simulated system, no such exchange is ever observed. Instead, their degree simply keeps growing through the process (not shown). However, Carl's as we said above, we must be aware about the relevance of this feature, since Carl's corpora do not exhibit so clear patterns of change in this way, maybe due to the fact that the child had a bit higher degree of maturation.

5. Discussion

Our study reveals two clearly differentiated behaviors in the early stages of language acquisition. Rules governing both grammatical and global behavior seem to be qualitatively and quantitatively different. Could we explain the transition in terms of self-organizing or purely external-driven mechanism? Clearly not, given the spe-

cial features exhibited by our evolving webs, not shared by *any* current model of evolving networks [10, 11].

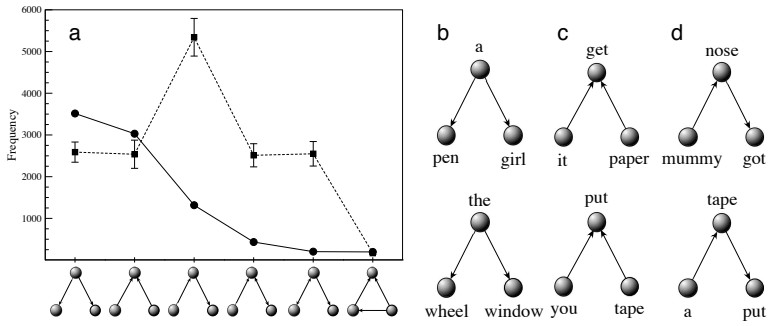


Fig. 9. Subgraph census plot for both real (circles) and simulated (squares) networks. As we can see in (a), there exist an overabundance of the first two subgraphs due to grammatical restrictions on the role of the syntactic head (see text). (b) and (c) are an example of the kind of nodes that participate in such small subgraphs. Beyond this two subgraphs, we find a sharp decay in its abundance against, compared with the model. This is due to the fact that the third studied motif (d) should be abundant (as in the model).

Beyond the transition, some features diverge dramatically from the pre transition graph. Such features cannot be explained from external factors (such as communication constraints among individuals). Instead, it seems tied to changes in the internal machinery of grammar. The sharp transition from small tree-like graphs to much larger scale-free nets, and the sudden change of the nature of hubs are the footprints of the emergence of new, powerful rules of exploration of the combinatorial space, i.e., the emergence of full adult syntax. This seems to support the hypotheses suggested by Hauser et al. [19]; see also [34]. The deviations in the role of hubs in the pre-transition stages observed between the two set of corpora could indicate that there exist different maturational speeds. This should imply that, despite Carl is able to produce functional words from the very begining, its role as articualtory elements of grammar is not accomplished until several months latter. At the other hand, Peter seems to introduce the functional particles when its cognitive system is ready to use them as the backbones of grammar.

Furthermore, we have presented a novel approach to language acquisition based on a simple, data-driven model. Previous model approaches based on self-organization cannot reproduce the observed patterns of change displayed by syntax graphs. Our main goal was to explore the potential roles of adaptive versus non-adaptive components in shaping syntax networks as they change in time. The model is able to reproduce some fundamental traits. Specifically we find that: (a) the global architecture of syntactic nets obtained during the acquisition process

can be reproduced by using a combination of Zipf's law and assuming a growing working memory and (b) strong deviations are observed when looking at the behavior of hubs and the distribution of subgraph abundances. Such disagreements cannot be fixed by additional rules. Instead, they indicate the presence of some innate, hard-wired component related with the combinatorial power of the underlying grammatical rules that is triggered at some point of the child's cognitive development. Our study supports the view that the topological organization of syntactic networks is a spandrel, a byproduct of communication and neural constraints. But the marked differences found here cannot be reduced to such scenario and need to be of adaptive nature. Furthermore, our analysis provides a quantitative argument to go forward beyond statistics in the search of fundamental rules of syntax, as it was early argued in [30].

A further line of research should extend the analysis to other (typologically different) languages and clarify the nature of the innovation. Preliminary work using three different european languages supports our previous results (Corominas-Murtra et al *unpublished work*). Moreover, modeling the transitions from finite grammars to unbounded ones by means of connectionist approximations [46] could shed light on the neuronal prerequisites canalizing the acquisition process towards a fully developed grammar as described and measured by our network approach.

6. Acknowledgements

The authors thank Guy Montag and the members of the CSL for useful discussions and Liliana Tolchinsky and Joana Rosselló for helpful comments. Also to Maria Farriols i Valldaura for her support during the whole process of this work. This work has been supported by grants IST-FET ECAGENTS under EU R&D contract 01194, the McDonnell foundation (RVS) and by the Santa Fe Institute.

99

References

[1] Barabási, A.-L. and Albert, R., Emergence of scaling in random networks, *Science* **286** (1999) 509.
[2] Bickerton, D., *Language and Species* (University of Chicago Press. Chicago, 1990).
[3] Bloom, L., Hood, L., and Lightbown, P., Imitation in language development if when and why, *Cognitive Psychology* (1974) 380–420.
[4] Bloom, L., Lightbown, P., and Hood, L., Structure and variation in child language, *Monographs of the society for Research in Child Development. Serial 160* (1975).
[5] Bollobás, B., *Random Graphs* (Cambridge University Press, 2001).
[6] Chomsky, N., *Language and problems of knowledge* (MIT Press. Cambridge, Mass, 1988).
[7] Christiansen, M. H. and Chater, N., Language as shaped by the brain, *Behavioral and Brain Sciences* **31** (2008) 489–509.
[8] Christiansen, M. H. and Kirby, S., Language evolution: Consensus and controversies, *Trends in Cognitive Sciences* **7** (2003) 300–307.

[9] Corominas-Murtra, B., Network statistics on early english syntax: Structural criteria, *arXiv.org:0704.3708* (2007).

[10] Dorogovtsev, S. N. and Mendes, J. F. F., Language as an evolving word web, *Proc. Royal Soc. London B* **268** (2001).

[11] Dorogovtsev, S. N. and Mendes, J. F. F., *Evolution of Networks* (Oxford University Press. New York, 2003).

[12] Elman, J. L., Learning and development in neural networks: The importance of starting small, *Cognition* **48** (1993) 71–99.

[13] Erdös, P. and Rényi, A., On random graphs, *Publicationes Mathemticae (Debrecen)* **6** (1959) 290–297.

[14] Ferrer-i-Cancho, R., Köhler, R., and Solé, R. V., Patterns in syntactic dependency networks, *Phys. Rev. E* **69** (2004) 051915.

[15] Ferrer-i-Cancho, R., Riordan, O., and Bollobás, B., The consequences of zipf's law for syntax and symbolic reference, *Proceedings of The Royal Society of London. Series B, Biological Sciences* (2005).

[16] Ferrer-i-Cancho, R. and Solé, R. V., The small world of human language, *Proc. Royal Soc. London B* **268** (2001).

[17] Ferrer-i-Cancho, R. and Solé, R. V., Least effort and the origins of scaling in human language, *Proc. Natl. Acad. Sci. USA* **100** (2003) 788–791.

[18] Harremoës, P. and Topsoe, F., Maximum entropy fundamentals, *Entropy* **3** [**3**] (2001) 191–226.

[19] Hauser, M. D., Chomsky, N., and Fitch, T. W., The faculty of language: What is it, who has it, and how did it evolve?, *Science* **298** (2002) 1569–1579.

[20] Holland, P. W. and Leinhardt, S., A method for detecting structure in sociometric data, *Am. J. of Soc.* **70** (1970) 492–513.

[21] Hudson, R., *Language Networks: The New Word Grammar* (Oxford University Press. New York, 2006).

[22] Hurford, J., Biological evolution of the saussurean sign as a component of the language acquisition device, *Lingua* **77** (1989) 187–222.

[23] Kauffman, S. A., *The Origins of Order: Self-Organization and Selection in Evolution* (Oxford University Press, 1993).

[24] Ke, J., Complex networks and human language, *arXiv:cs/0701135* (2007).

[25] Komarova, N., Niyogi, P., and Nowak, M., The evolutionary dynamics of grammar acquisition, *J. Theor. Biol.* **209** (2001) 43–59.

[26] Komarova, N. L. and Niyogi, P., Optimizing the mutual intelligibility of linguistic agents in a shared world, *Art. Int.* **154** (2004) 1–42.

[27] Macwhinney, B., The emergence of linguistic form in time, *Connection Science* **17** (2005) 191–211.

[28] Maynard-Smith, J. and Szathmàry, E., *The Major Transitions in Evolution* (University of New York Press. New York, 1997).

[29] Melçuck, I., *Dependency Grammar: Theory and Practice* (Oxford University Press. New York, 1989).

[30] Miller, G. A. and Chomsky, N., Finitary models of language users, in *Handbook of Mathematical Psychology*, eds. Luce, R. D., Bush, R., and Galanter, E., Vol. 2 (Wiley, New York, 1963), pp. 419–491.

[31] Newman, M. E. J., Assortative mixing in networks, *Phys. Rev. Lett.* **89** (2002) 208701.

[32] Newport, E. L., Maturational constraints on language learning., *Cogn. Sci.* **14** (1990) 11–28.

[33] Niyogi, P., *The Computational Nature of Language Learning and Evolution* (MIT Press. Cambridge, Mass., 2006).

[34] Nowak, M. A. and Krakauer, D., The evolution of language, *Proc. Nat. Acad. Sci. USA* **96** (1999) 8028–8033.

[35] Oudeyer, P.-Y., *Self-Organization in the Evolution of Speech*, Studies in the Evolution of Language (Oxford University Press, 2006).

[36] Pinker, S., *The Language Instinct* (Penguin Books: London, 1994).

[37] Pinker, S. and Bloom, P., Natural language and natural selection, *Behav. Brain Sc.* (1990) 707–786.

[38] Radford, A., *Syntactic Theory and the Acquisition of English Syntax: the nature of early child grammars of English* (Oxford. Blackwell, 1990).

[39] Sigman, M. and Cecchi, G., Global organization of the wordnet lexicon, *Proc. Nat. Acad. Sci. USA* **99** (2002) 1742–1747.

[40] Solé, R. and Goodwin, B., *Signs of Life: How Complexity Pervades Biology.* (New York, NY. Basic Books, 2001).

[41] Solé, R. V., Syntax for free?, *Nature* **434** (2005) 289.

[42] Steels, L., The synthetic modelling of language origins, *Evolution of Communication* **1(1)** (1997) 1–34.

[43] Steels, L., Evolving grounded communication for robots, *Trends in Cognitive Sciences* **7** (2003) 308–312.

[44] Steyvers, M. and Tenenbaum, J. B., The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth, *Cogn. Sci.* **29** (2005) 41–78.

[45] Szamadó, S. and Szathmàry, E., Selective scenarios for the emergence of natural language, *Trends in Ecology and Evolution* **21** (2006) 555–61.

[46] Szathmáry, E., Szatmáry, Z., Ittzés, P., Orbán, G., Zachár, I., Huszár, F., Fedor, A., Varga, M., and Számadó, S., In silico evolutionary developmental neurobiology and the origin of natural language., In: Lyon, C. and Nehaniv, C. L. and Cangelosi, A. *Emergence of Communication and Language.* Springer-Verlag. London pp. 151 – 187 (2007).

[47] Theakston, A. L., Lieven, E. V. M., Pine, J. M., and Rowland, C. F., The role of performance limitations in the acquisition of verb-argument structure: an alternative account., *Journal of Child Language* **28** (2001) 127–152.

[48] Veneziano, E. and Sinclair, H., The changing status of "filler syllables" on the way to grammatical morphemes, *Journal of Child Language* **27** (2000) 461–500.

[49] Wasserman, S. and Faust, K., *Social Network Analysis* (Cambridge. Cambridge University Press, 1994).

[50] Watts, D. J. and Strogatz, S. H., Collective dynamics of 'small-world' networks., *Nature* **393** (1998) 440–442.

Coloring the Syntax Networks of Child Language

Bernat Corominas Murtra,[1] Martí Sànchez Fibla,[2] Sergi Valverde,[1] and Ricard Solé[1, 3, 4]
[1]*ICREA-Complex Systems Lab, Universitat Pompeu Fabra (GRIB), Dr Aiguader 80, 08003 Barcelona, Spain*
[2]*SPECS, Technology Department, Universitat Pompeu Fabra, Carrer de Roc Boronat 138, 08018 Barcelona, Spain.*
[3]*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM 87501, USA*
[4]*Institut de Biologia Evolutiva. CSIC-UPF. Passeig Maritim de la Barceloneta, 37-49, 08003 Barcelona, Spain*

The emergence of syntax during the childhood is a remarkable example of the emergence of complexity in the natural world. In this work we study two well-known case studies from the perspective of network theory but, as a difference with other approaches, we explore the evolution of syntax networks using a non-statistical but powerful indicator of complexity, the *chromatic number*. It is shown how the chromatic number is able to capture the huge emergence of complexity of the underlying syntax rules around the age of two. Furthermore, we compare the data against a null model of sentence production, obtaining interesting deviations of the model values from the real ones; deviations that are hardly identified by using standard statistical parameters of network description. In a more general level, we observe that the chromatic classes define independent regions of the graph, and thus, can be interpreted as the footprints of some incompatibility relations, an interpretation that sets the chromatic number as an opposed concept to modularity. We finally emphasize that although shifts in the chromatic number have been largely studied for model graphs at the theoretical level, this is, to our knowledge, the first identification of chromatic transitions in the evolution of a real system.

Keywords: Complex Networks, Graph Colouring, Modularity, Syntax

I. INTRODUCTION

The emergence of syntax both in early chilhood or at the philogenetic level is one of the most fascinating natural phenomenon (Bickerton, 1990; Christiansen and Kirby, 2003; Hauser *et al.*, 2002; Maynard-Smith and Szathmàry, 1997), being it a remarkable example of emergence of combinatorial complexity in natural world. Indeed, around the age of two, linguistic structures produced by children display a qualitative shift on their complexity, indicating a deep change on the rules underlying them (Corominas-Murtra *et al.*, 2009; Radford, 1990). This sudden increase of grammar complexity is known as the *syntactic spurt*, and draws a border between the *two words* stage, where only isolated words or combinations of two words occur, to a stage where the grammar rules governing this syntax are close to the one we can find in adult speech -although the cognitive maturation of kids makes the semantic content or the pronunciation different from the adult one.

The above described process is the history of the emergence of complex combinatorial patterns enabled and at the same time, constrained, by the underlying grammar rules. The combinatorial nature of syntactic relations naturally leads the network approach as a good statistical tool to explore this fundamental building block of grammar (Corominas-Murtra *et al.*, 2009; Ferrer-i-Cancho *et al.*, 2004; Ke, 2007; Solé *et al.*, 2010). In this network, words are nodes and links the projection of actual syntactic relations present in the studied set of data (Corominas-Murtra, 2007; Corominas-Murtra *et al.*, 2009). The aggregation of the words and the syntactic relations of a sufficiently larger linguistic corpora enables us to extract a global view of the use of grammar. In this way, the exploration of the different syntactic networks obtained by grouping all the syntactic projections of a single individual belonging to the same stage of the acquisition process has shown interesting features concerning the evolution of the complexity of the network, displaying an abrupt increase of the complexity indicators of the net at the age of two -in agreement to the above described *syntactic spurt*. Specifically, it has been observed that, from the very beginning, although many words appear isolate, there is a giant connected component -the maximal set of nodes by which for every pair of them there is a finite path- containing almost all syntactic relations, and thereby easily identifiable as the *seed* from which the net will grow during the acquisition process. This giant connected component (GCC, henceforth) experiences an abrupt increase at the age of two, when the syntactic spurt takes place. This sudden increase coincides with a the emergence of functional particles -mainly, determiners and prepositions- which provoke a complete reorganization of the net. Indeed, these functional words are seldom used at the two words stage, but just after the increase of the net, they set up as the *hubs* -the most connected nodes- of the net. This sudden growth, jointly with the fact that from the very beginning the system displays a clearly defined GCC, does not fit with standard percolation phenomena, and points to deeper explanations related to the internal endowment and growing cognitive apparatus of the child. It is worth to note that motif structure, clustering and average connectivity also display a shift during the syntactic spurt (Corominas-Murtra *et al.*, 2009).

As any approach, there are limitations, and they must to be pointed out. The first one is the nature of "words" as the building blocks of syntax, for it is clear that the

realization of grammatical functions within a sentence is differently distributed depending on the language. In this way, for example, the future tense in English is realized as a differentiated word, *will*, whereas in French and other romanic languages the future tense is realized as a small piece of a bigger word which is the verb in future tense. There is another constraint, often diminished: The fact that syntactic structures are not defined as binary word-word relations (as they appear in the graph representation), but instead they are defined as hierarchical relations among structured sets of words -or, more generally, grammatical elements. Therefore, a syntactic graph is actually a graph of *projected syntactic relations* into word-word relations. Does it invalidate the network approach to syntax? Absolutely not, for it provided for the first time an abstraction able to manage an arbitrary large set of linguistic productions as a whole, where the combinatorial role of the syntactic relations defines, although through projections where some information is lost, the existence of links. Moreover, as we shall see, early periods of syntax acquisition can be reproduced with a grammar having linear relations.

With all necessary caution, it is clear that the statistics of the network-like exploration of syntactic patterns provided a valuable source of new data to understand how language works and evolves at the ontogenetic level. However, there is something hardly to grasp using statistics, namely, some measure of the entanglement of the syntactic relations, which, even projected, strongly constrain the combinatorial possibilities of the syntactic graphs. We need thus stronger indicators of network complexity, going in depth into the footprints of compatibilities and incompatibilities imposed by the underlying grammar. In this study we propose to go further the statistics to explore the properties of a real network where it is known that internal relations of compatibility are at work. Specifically, in this work we study the evolution of the complexity of the syntactic networks through the acquisition process from two case studies by which the sequence of syntactic networks have been constructed in detail, using the powerful indicator of graph complexity provided by the *chromatic number* -and associated measures- of the graph (Bollobas, 1998; Bollobás, 2001; Brooks and Tutte, 1941). Roughly speaking, the chromatic number can be defined as the minimal number of *colors* needed to *paint* all nodes of the graph in such a way that no node is connected to a node having the same color. From the statistical physics we find an analogous problem, the so-called *Potts* model, in which the chromatic problem can be embedded, by assuming that $T \to 0$ (Wu, 1982). The q-coloring problem, i.e., to know wether a graph can be colored with q different colors went down in history as one of the most important NP-complete problems. However, the partition it defines among compatible sets of nodes offers a window into the complexity and constraints imposed over the combinatorial processes underlying the construction of the syntactic graph. Moreover, the observation of transitions along the syntactic acquisition process can provide a very valuable information over the overall pattern of syntactic production. It is worth to note that transitions in the evolution of the chromatic number have been widely studied at the mathematical level (Achlioptas and Molloy, 1999; Bollobás, 1988; Bollobás, 2001; Zdeborová and Krzakala, 2007). In this way, several transitions have been defined over a random graph of increasing connectivity. The nature of such transitions is sharp and revealed full of very intriguing phenomena. The exploration we made over sequences of syntactic graphs also displayed transitions in the chromatic number, which is, to the best of our knowledge, the first time that such transitions have been observed in real-world phenomena. We finally note that the information provided by the chromatic number is, fundamentally, an indicator of compatibility/incompatibility relations underlying graph structures. In this approach, classes of nodes would be defined precisely by the fact that there are no connections among them, a measure conceptually opposite to graph modularity.

The remaining of the paper is organized as follows. Section II is devoted to the definition of the object of study -a sequence of syntactic networks obtained during the acquisition process- and the theoretical tools to explore it. This latter point basically introduces the so-called *Potts model* as the way to introduce the chromatic number. In section III we apply these theoretical constructs to our problem and we analyze the obtained data by using different estimators of relevance, the most prominent of them being a null model of random sentence generation. In section IV we discuss the obtained results and we highlight the potential impact of this kind of complexity estimators on the field of complex networks.

II. SYNTAX GRAPHS AND COLORING: BASICS

This section is devoted to the presentation of the core concepts related to graph coloring. Beyond basic graph definitions, we present the coloring problem from an energetical point of view, by defining a Hamiltonian associated to both the graph and the coloring sequence which must be minimized. Such a minimization is a NP-complete problem and the theoretical steps of the algorithm are described according to the theoretical basis provided (Wu, 1982).

A. Graph Definitions

A graph $\mathcal{G}(V, E)$ -hereafter, $\mathcal{G}$- is composed by the set of $V = \{v_1, ..., v_n\}$ *nodes* and $E = \{e_1, ..., e_m\}$ *links*, which are pairs of different nodes $e_j = \{v_i, v_k\}$, being thus $E \subseteq V \times V$. Such pairs depict links among nodes and we assume they are unordered, since we are working with undirected graphs. The number of links attaching node v_i, to be noted $k(v_i)$ is the *degree* of the node v_i and $\langle k \rangle$ is the *average degree* of $\mathcal{G}$. The *degree distribu-*

tion, to be named $P(k)$ is the probability distribuition which accounts for the probability to select at random a node whose degree is k. The identity card of a graph is the so-called *Adjacency matrix*, $\mathbf{a}(\mathcal{G})$, which is defined as follows:

$$a_{ij} = \begin{cases} 1, \text{ iff } (\exists e_k \in E) : (e_k = \{v_i, v_j\}) \\ 0, \text{ Otherwise.} \end{cases} \quad (1)$$

In undirected graphs, such matrix is symmetrical.

B. Building the Networks of Early Syntax

Language acquisition process has been already studied in the past from a network perspective. Different types of networks have been defined, obtaining from them interesting, complementary results, providing a new view of the evolution of global patterns of language during the acquisition process.

We will follow the definition of syntactic network provided in (Corominas-Murtra, 2007; Corominas-Murtra *et al.*, 2009) -see also (Solé *et al.*, 2010). In these approaches, syntactic networks are built *by hand* -using the software provided in (Hristea and Popescu, 2003)- from a given corpora by projecting constituent structure -the basic phrase structure of a linguistic production- of children's utterances into linear relations among lexical items, and then, aggregating all the productions in a graph (Corominas-Murtra, 2007; Corominas-Murtra *et al.*, 2009)[1].

The two case studies are obtained from the well-known CHILDES Database (Bloom *et al.*, 1974, 1975). Specifically, we choose Peter and Carl's corpora, a sequence of recorded conversations where the child is not conditioned at all. For both Peter and Carl's corpora, we choose 11 different recorded conversations distributed in approximately uniform time intervals from the age of $\sim$ 20 months to the age of $\sim$ 28 months, the period in which the syntactic spurt takes place. From every recorded conversation, we extract the syntactic network of child's utterances obtaining a sequence of 11 syntactic graphs corresponding to the sequence of Peter conversations $\mathcal{G}_{P1}, ..., \mathcal{G}_{P11}$ and Carl's conversations $\mathcal{G}_{C1}, ..., \mathcal{G}_{C11}$.

C. The Potts Model and the Chromatic Problem

The chromatic problem can be formulated in the following terms: *What is the minimal number of 'colours' needed to paint all nodes of the graph in such a way that no node is connected to other nodes of the same colour?*. Such a combinatorial problem can be handled -and algorithmically implemented- from a theoretical physics viewpoint by observing the close relation it has with the Antiferromagnetic q-dimensional Potts model at $T = 0$ (Wu, 1982). This model is a generalization of the classical Ising model for lattices, where in every node there is a particle displaying a spin which energetically constraints the state of its neighbors. Spins have only two states, namely $|\uparrow\rangle$ and $|\downarrow\rangle$, and the Potts model offers a generalization where compatibility relations take into account an arbitrary number q of different states.

The physical formulation of the problem can be stated as follows. Let us consider a partition of the set of nodes V containing q different classes, namely, $G_q(V) = \{g_1, ..., g_q\}$ of V, i.e.:

$$\bigcap G_q = \varnothing \text{ and } \bigcup G_q = V, \quad (2)$$

The *state* of node v_i, σ_i, will be given by the class of $G_q(V)$ to which v_i belongs, i.e., $\sigma_i \in g_j$. Let $\mathcal{F}_q(V)$ be the ensemble of all partitions of V containing q different classes. Every element of $\mathcal{F}_q(V)$ will have an associated energy penalty depicted by the following hamiltonian:

$$\mathcal{H}(G_q) = J \sum_{i<j} a_{ij} \delta(\sigma_i, \sigma_j), \quad (3)$$

where J is the *coupling constant* -which we set to $J = 1$, for the sake of simplicity- and δ the Kronecker symbol, defined as [2]

$$\delta(\sigma_i, \sigma_j) = \begin{cases} 1, \text{ iff } i = j \\ 0, \text{ Otherwise.} \end{cases} \quad (4)$$

It is clear that, the higher the presence of pairs of connected nodes belonging to the same state, the higher will be the energy of the global state of the graph. Given a fixed q, the configurations displaying minimal energy may have an amount of non-solvable situations, leading to the unavoidable presence of connected nodes at the same state. This phenomenon is called *frustration*, and for these configurations, the ground state of the Hamiltonian defined in (3) displays positive energy. In general, if the case is not the one described above, i.e., $(\exists G_q \in \mathcal{F}_q(V))$ by which

$$\mathcal{H}(G_q) = 0, \quad (5)$$

we say that the graph is q-colorable, being the q different *colors* the q different classes which are the members of G_q. We observe that, if the graph is q-colorable, there exist at least one partition $G_q \in \mathcal{F}_q(V)$ such that, if $v_i, v_j \in V$ belong to the same *class* or *color* of the partition, namely $g_l \in G_q$, then:

$$(v_i, v_j \in g_l) \Rightarrow a_{ij} = 0. \quad (6)$$

[1] The subtleties of child's language, including the presence of imitations or the fuzzy definition of structures implies an additional difficulty to the definition of the network. The detailed discussion is provided in (Corominas-Murtra, 2007)

[2] the energy units of this Hamiltonian are arbitrary.

Relation (6) makes explicit that if a graph is successfully colored with q colors (i.e., we have a partition such that $\mathcal{H}(G_q) = 0$), color classes are sets of disjoint elements of the graph. The coloring problem consist in finding the minimal number of classes or colors we need to properly *paint* the graph, which is the so-called *Chromatic Number* of the graph $\mathcal{G}$, notated $\chi(\mathcal{G})$. In formal terms, given a graph $\mathcal{G}$, we are looking for the following number:

$$\chi(\mathcal{G}) = \inf\{q : (\exists G_q \in \mathcal{F}_q(V)) : \mathcal{H}(G_q) = 0\}. \quad (7)$$

The search for the chromatic number of a given arbitrary graph is known to be one of the most famous NP-complete problems.

Despite the high complexity of the problem, several bounds can be defined. We begin with a lower bound, the so-called *Clique* number, $\omega(\mathcal{G})$, which is the number of nodes belonging to the largest subgraph of $\mathcal{G}$ in which every node is connected to all other nodes of the subgraph -i.e, the largest *clique* of the graph. Finding the clique number is also a NP-complete problem. The interest on this quantity lies on the fact that it defines a natural lower bound, for $\chi(\mathcal{G})$,

$$\omega(\mathcal{G}) \leq \chi(\mathcal{G}). \quad (8)$$

Alternatively, an upper bound on $\chi(\mathcal{G})$ can be defined by looking at the K-core structure of $\mathcal{G}$. The $K(\mathcal{G})$ is the largest subgraph whose nodes display degree higher or equal to K. Let $K^*(\mathcal{G})$ be defined as:

$$K^* = \max\{K : K(\mathcal{G}) \neq \varnothing\}, \quad (9)$$

i.e., the K-core with largest connectivity that can be found in $\mathcal{G}$. Then, it can be shown that (Bollobás: 1999)

$$\chi(\mathcal{G}) \leq K^* + 1. \quad (10)$$

For some families of random graphs, however, it has been found that the chromatic number has an asymptotic behavior depending on the connectivity, following a relation of the form $\chi(\mathcal{G}) \sim \frac{\langle k \rangle}{\log\langle k \rangle}$ (Bollobás, 2001). For scale-free networks whose exponent γ lies in the interval $2 < \gamma < 3$, however, the divergence of the clique number with the size of the graph, even at constant average connectivity, sweeps away the chromatic number -see eq. (8)- diverging with graph's size too (Bianconi and Marsili, 2006).

D. Colorability and Optimization

The q-coloring problem, i.e., finding if the graph $\mathcal{G}(V, E)$ admits a proper node coloring (one where to adjacent nodes have a different color), with q colors is NP-complete for $q > 2$. Computing the *Chromatic Number* is an NP-complete problem and can be reduced to a sequence of q-coloring problems. The sequence of q-coloring problems consists in searching the partition(s)

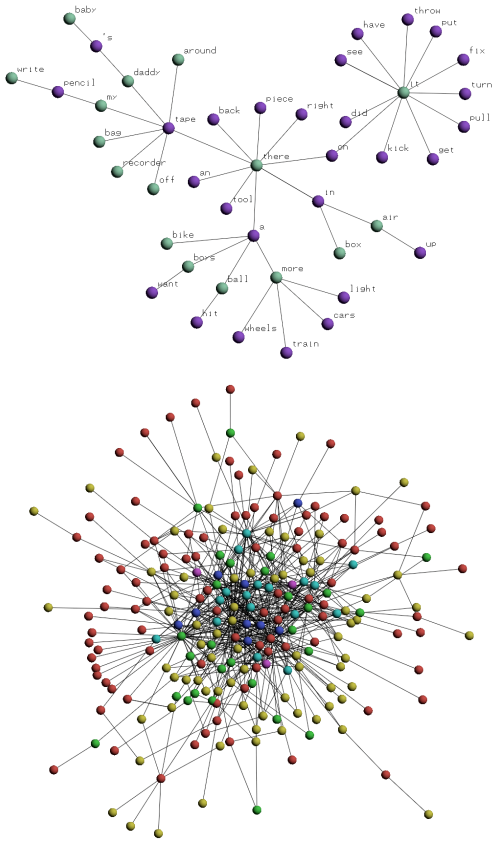


FIG. 1 Evolution of ω, χ and $K^* + 1$ during the acquisition process. Above, we find the chromatic number obtained from the real networks.

$G_q^* \in \mathcal{F}_q(V)$ which minimize the energy function defined in eq. (3), i.e,

$$G_q^* = \min_{G_q \in \mathcal{F}_q(V)} \{\mathcal{H}(G_q)\}. \quad (11)$$

Starting from 1-coloring and increasing k until a feasible coloring is found, i.e., we reach the value k provided in eq. (7) . In general, we will have a decreasing sequence of energies ending at $\mathcal{H}(G_q^*) = 0$, namely:

$$\mathcal{H}(G_1^*) \leq ... \leq \mathcal{H}(G_q^*) = 0, \quad (12)$$

where q satisfies eq. (7). A usual formulation of q-coloring in terms of a Constraint Satisfaction Problem (CSP), has one variable per node, each variable taking q possible values. A "value-different" constraint exists between two variables if the corresponding nodes are linked by an edge in the original graph, obliging the variables to take a different value and thus a different color in a solution. The *Chromatic Number* can be computed solving

the sequence of q-coloring problems formulated as CSP. To additionally compute the sequence of energy values associated to a given q-coloring, we solve the corresponding NP-complete problem of minimizing the number of edge violations. For this purpose we need to trivially extend our formulation of a CSP to a Weighted CSP problem (WCSP). Each edge violation will have a cost of one and a graph accepts a valid q-coloring if the minimal total cost is 0. The decreasing sequence of edge violations are shown in table I.

As we pointed out above, finding the maximum clique is also an NP-complete problem. We use here an Integer Linear Programming (ILP) formulation [3]. Similarly to the previous CSP formulation, we define a variable per node, each variable taking 2 possible values $x_i = \{0, 1\}$ whether node i belongs to the maximum clique (value 0) or not (value 1). The ILP formulation maximizes the number of nodes assigned to the clique by minimizing $min \sum_{i=1}^{n} x_i$ such that $(1-x_i)+(1-x_j) \leq 1, \forall (i,j) \in \overline{E}$. These latter constraint forbids every two nodes that don't have a link to belong at the same time to the maximum clique. To be able to express the fact that two nodes don't have a link we use the complementary of the edge set that we denote $\overline{E}$.

III. THE EVOLUTION OF χ IN SYNTAX ACQUISITION

In this section we study the evolution of the chromatic number of the sequence of syntactic networks obtained during the acquisition stage. Results are validated by computing the bounds provided by the clique number and the value of the K-core having largest connectivity. Furthermore, acknowledging the non-statistical nature of the chromatic number, we provide a table of relevance of the actual chromatic numbers against the minimum energies of those colorings by which there not exist a configuration $G_q \in \mathcal{F}_q$ such that $\mathcal{H}(G_q) = 0$. This latter estimator and the bounds we discussed above enable us to identify whether the chromatic number can be attributed to a global complexity pattern of the nets or, on the contrary, its origin lies on an anomalous behavior of a small region of the net. The validation ends by comparing the obtained data against the nets obtained through a random sentence generator described in sec. III.B.

A. Evaluating Data

What is the evolution of the Chromatic number through the process of acquisition of syntax? To answer this question, we computed the sequence of energies for optimal configurations from $q = 1$ to $\chi(\mathcal{G})$ -see eqs. (3,7,12)- for all syntactic graphs corresponding to the child's utterances, namely $\mathcal{G}_{P1}, ..., \mathcal{G}_{P11}$ and $\mathcal{G}_{C1}, .., \mathcal{G}_{C11}$ -see sec. II.B- thus obtaining two sequences of chromatic numbers, one corresponding to the evolution colorability of Peter's syntactic graphs, $s_P(\chi)$, and the other corresponding to the evolution of colorability of Carl's syntactic graphs, $s_C(\chi)$:

$$\begin{aligned} s_P(\chi) &= \chi(\mathcal{G}_{P1}), ..., \chi(\mathcal{G}_{P11}) \\ s_C(\chi) &= \chi(\mathcal{G}_{C1}), ..., \chi(\mathcal{G}_{C11}). \end{aligned}$$

These sequences are our main object of study.

We defined different controls to evaluate the relevance of the chromatic number. The reason to define such controls stems come from the fact that $\chi(\mathcal{G})$ is not a statistical parameter and its behavior can be biased by the strange behavior of small parts of the graph. Thus, we also computed the clique number $\omega(\mathcal{G})$ and the connectivity of the most connected K-core, K^*, defined in eq. (9) thus exploring the behavior of the natural lower and upper bounds on the colorability -see eq. (8) and eq. (10). Therefore, every sequence $s_P(\chi), s_C(\chi)$ will be accompanied by two sequences, namely Ω, κ:

$$\begin{aligned} \Omega_{P,C} &= \omega(\mathcal{G}_{P1,C1}), ..., \omega(\mathcal{G}_{P11,C11}) \\ \kappa_{P,C} &= K^*(\mathcal{G}_{P1,C1}), ..., K^*(\mathcal{G}_{P11,C11}). \end{aligned}$$

Furthermore, if a network has a chromatic index equal to $\chi(\mathcal{G}) = q$ we evaluate the relative impact of the energy of the optimal configuration(s) having $q-1$ different colors. Such impact is evaluated as the minimal relative number of *frustrated* vertices:

$$f_{q-1}(\chi) = \frac{\mathcal{H}(G^*_{q-1})}{|E|}. \tag{13}$$

Analogously, we can define $f_{q-2}(\chi)$, etc... until $q-i=1$, where, by definition, $f_{q-i}(\chi) = 1$. $f_k(\chi)$ is an estimator of the relevance of the q-coloring which, jointly to $\omega(\mathcal{G})$ and K^* enables us to evaluate wether the chromatic index is due to the behavior of a small number of nodes or if it is the natural outcome given the global features of the graph. Since every graph has an associated measure of the relevance of χ, we will have a sequence of f's for $s_P(\chi)$ and another associated to $s_C(\chi)$, namely $f_P(\chi)$ and $f_C(\chi)$.

Both the sequences $s_C(\chi)$ and $s_P(\chi)$ display close behaviors, although interesting differences can be observed -see figs. (2, 4). The first corpora displays constant, low values of the chromatic number (see table for relevance terms) and, at the stage when the syntactic spurt takes place, the chromatic index displays a transition to higher values. Particularly interesting is the bipartite nature of the three first stages of s_P. It is worth to note that the 4th network has $\chi(\mathcal{G}) = 3$, but, as we can observe both in the evolution of ω -see fig. (2) - and in the table of relevance of the chromatic numbers -see table I-, this jump from 2 to 3 is due to the presence of a single triangle in a

[3] We translate the ILP formulation into a WCSP minimization problem because we have solved all problem formulations, including the q-coloring instances, with an optimization library, toulbar2, that by design can only minimize.

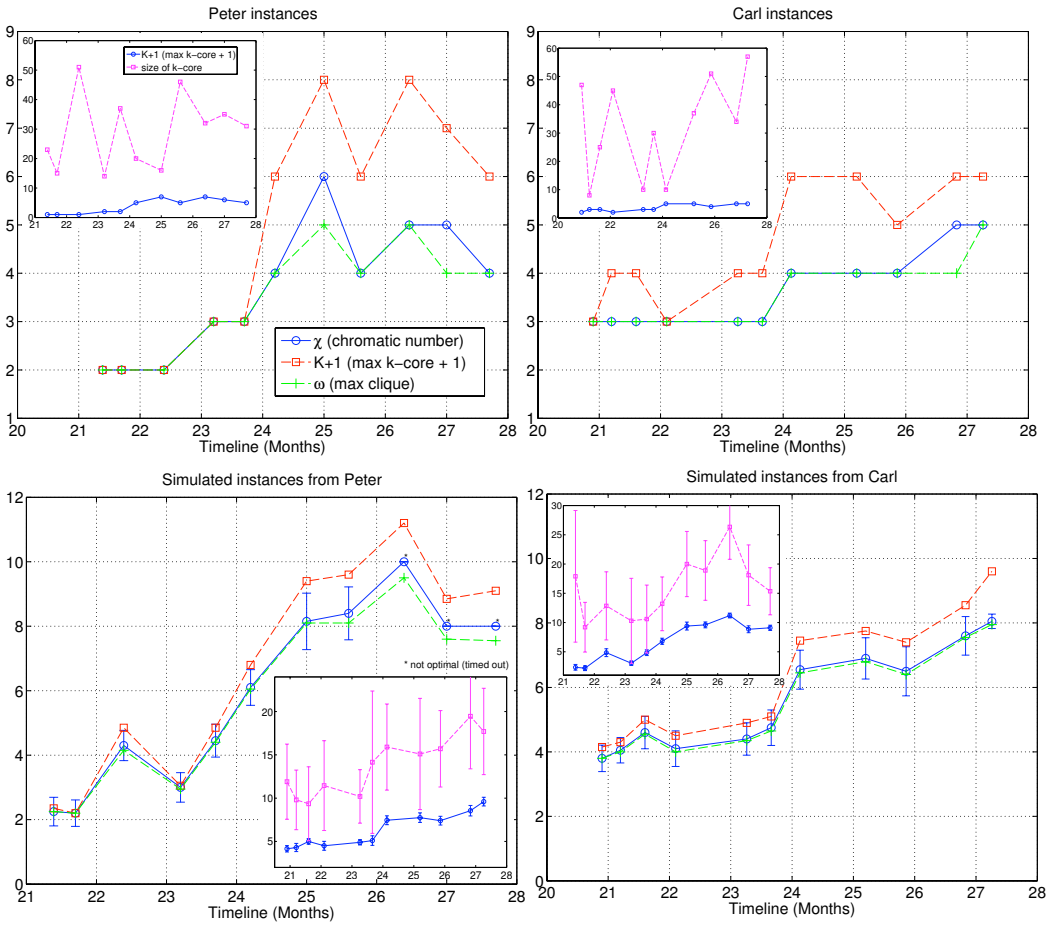


FIG. 2 Evolution of ω, χ and $K^* + 1$ during the acquisition process. Above, we find the chromatic number obtained from the real networks. Consistently with the emergence of complex syntax rules, both plots display a huge increasing trend on the chromatic number during the studied period, being this trend sharper around the age of two. Particularly interesting is the case of Peter, where the three first networks are bipartite, an identity card of the so-called ”two-word stage” -see text. We observe that the chromatic number equals the clique number in most cases, suggesting that the presence of a clique is the responsible of the value of this index, an intuition that can be validated using the values of table (I). The plots on the bottom display the evolution of ω, χ and $K^* + 1$ using the random sentence generator described in section III.B. INterestingly, even the global trend is close to the one observed in real nets, all complexity indicators display clearly higher values. This points to the action of underlying grammar, which would, at the same time, be responsible of the emergence of combinatorial complexity but also a constraint to it.

bipartite network. Indeed, this highly restrictive feature can be straightforwardly identified as the foot-prints of the so-called 2-word stage, in which syntactic structures have at least 2 elements and where the kind of relations among the building blocks of such productions is strongly constrained by their semantic content. Roughly speaking, the grammar at this stage mainly generates pairs of complementary words, like

$$\langle \text{verb}, \text{noun} \rangle \text{ or}$$

$$\langle \text{adjective}, \text{noun} \rangle;$$

(e.g ”car red” or ” *horsie* run”). This grammar is highly restrictive -structures like $\langle \text{verb}, \text{verb} \rangle$ do not exist, for example. This lack of flexibility is due to the absence of functional particles -such as ”a”, or ”the”. It is not the case of the sequence belonging to Carl, s_C, where from the very beginning $\chi \geq 3$, and it is worth to note the presence of functional particles from the very beginning, making the grammar more flexible and thereby generat-

	$\mathcal{G}_{P1}$	$\mathcal{G}_{P2}$	$\mathcal{G}_{P3}$	$\mathcal{G}_{P4}$	$\mathcal{G}_{P5}$	$\mathcal{G}_{P6}$	$\mathcal{G}_{P7}$	$\mathcal{G}_{P8}$	$\mathcal{G}_{P9}$	$\mathcal{G}_{P10}$	$\mathcal{G}_{P11}$
$f_1(\chi)$	1	1	1	1	1	1	1	1	1	1	1
$f_2(\chi)$	0	0	0	1/49	5/105	66/434	131/644	87/589	157/903	104/659	95/717
$f_3(\chi)$	0	0	0	0	0	8/434	31/644	15/589	40/903	20/659	10/717
$f_4(\chi)$	0	0	0	0	0	0	8/644	0	8/903	2/659	0
$f_5(\chi)$	0	0	0	0	0	0	1/644	0	0	0	0
	$\mathcal{G}_{C1}$	$\mathcal{G}_{C2}$	$\mathcal{G}_{C3}$	$\mathcal{G}_{C4}$	$\mathcal{G}_{C5}$	$\mathcal{G}_{C6}$	$\mathcal{G}_{C7}$	$\mathcal{G}_{C8}$	$\mathcal{G}_{C9}$	$\mathcal{G}_{C10}$	$\mathcal{G}_{C11}$
$f_1(\chi)$	1	1	1	1	1	1	1	1	1	1	1
$f_2(\chi)$	6/140	5/119	11/156	6/128	10/152	14/199	61/361	65/442	71/439	93/592	131/687
$f_3(\chi)$	0	0	0	0	0	0	9/361	11/442	8/439	16/592	29/687
$f_4(\chi)$	0	0	0	0	0	0	0	0	0	1/592	4/687

TABLE I This table shows the relative values of the energies associated to a given $q-$coloring. The definition of the function $f_i(\chi)$ -see eq. (13)- enables us to evaluate the statistical significance of the chromatic index, which is non-statistical in nature.

ing higher chromatic numbers.

The middle stages of both collection of corpora display an increase on the chromatic number, being sharper in the case of s_P, which is fully consistent with the process of the emergence of a complex syntax. Both the table of relevance -see table I-, the values of the largest clique and the size of the maximum K-core -see fig. (2)-suggest that the final chromatic number is due, in general, to the presence of cliques of higher order, but if we go to f_{q-1} we observe that the chromatic number can no longer be attributed to a single clique; its relevance as a global complexity estimator is much more feasible. Furthermore, both sequences of corpora display a divergence of K^* from both the chromatic number and the clique number, which tells us that the whole structure of the net -or an important part of it- has enough connectivity to enable the emergence of a maximum K-core which is not a trivial clique. This latter feature is reinforced when looking at the sizes of the maximum K-cores -see fig. (2)-, which are, in general, more than twice the number of nodes contained in the maximum clique. Thus, even high values of K^* can quickly lead to a K-core containing a large number of nodes, which means that grammar generates a collection of compatibility relations able to generate a great amount of combinatorial complexity.

B. Evolution of χ in a model of random sentence generation

The process of data validation ends with the comparison of real data against a null model, in which no syntax is at work. This data-driven model, described in detail in (Corominas-Murtra *et al.*, 2009), generates a randomized version of child's utterances, by designing an *engine* which sends strings of lexical items respecting the statistics of real data, except in the fact that structures are defined at random. This model has been run extracting the data from the 11 recorded conversations belonging to both Peter and Carl's corpora. It is, thus, a null model of utterance production, not a null model of network. There are strong reasons to consider this null model validation more suitable than a more straightforward one, e.g. considering a standard algorithm of network randomization. The most salient one is that our object of study is the syntax of children's productions, and the network abstraction is our way to organize data. Therefore, the null model is not a graph-like null model but a syntax free model of sentence generation upon which the graph is built.

The statistical parameters to generate the ensemble of null models of production is based on:

1. The number of sentences produced by the child in the studied corpora $|S_P(i)|$, $S_C(i)$, for Peter and Carl's corpora, respectively.

2. The *structure length*: the structure length of of the i-th structured production (s_i) of a given corpora is measured by counting the number of words that participate in the i-th syntactic structure. We collect this data for all two collections of 11 corpora and we obtain the probability distribution that accounts for the probability that a given structure contains s words in a given corpus.

3. we assume that the probability of the i-th most frequent word is a scaling law:

$$p(i) = \frac{1}{Z} i^{-\beta} \tag{14}$$

with $1 \leq i \leq N_w(T)$, $\beta \approx 1$ (i.e., this distribution is the so-called Zipf's law) and Z is the normalization constant:

$$Z = \sum_{i=1}^{N_w(T)} \left(\frac{1}{i}\right)^{\beta} \tag{15}$$

(notice that Z depends on lexicon size, $N_w(T)$, which grows slowly at this stage). The assumption the frequency of words follows the so-called Zipf's law is widely supported by empirical data.

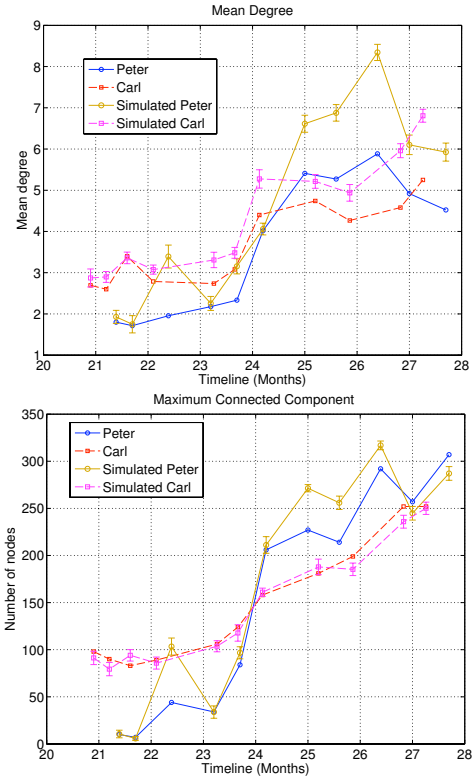


FIG. 3 Evolution of the average connectivity (above) and size of the connected component (below) of both real nets and the ensemble generated by the random sentence generator. As shown in (Corominas-Murtra *et al.*, 2009) these statistical indicators display a huge increase during the studied period, being this increase sharper around the age of two, i.e., during the syntactic spurt. It is worth to note that this random generator is able to generate networks whose size and connectivity is close to the real ones. As discussed in section II.C both the mean connectivity and the size of the net play a crucial important role in determining the values of ω, χ and K^*.

Once we extract this information from the specific corpus under study, we run this model by generating $|S_{P,C}(i)|$ random sentences constrained by the statistical parameters 2) and 3) of the corpus. For every corpus we generate an ensemble of 20 replicas.

The computation of the chromatic number generates the sequences $\tilde{s}_P(\chi), \tilde{s}_C(\chi)$, which is the sequence of the average chromatic numbers over the graph ensemble of randomized versions. Analogously, we generate the sequences $\tilde{\Omega}_{P,C}$ and $\tilde{\kappa}_{P,C}$ of the average clique numbers and maximum K-cores, respectively.

The most salient property we find when comparing real networks obtained from both Peter and Carl's corpora with their randomized counterparts is the huge increase of χ, ω and K^* we find in the random versions. Indeed, the ensemble of random strings, even sharing the same statistics of the real data, displays a huge increase in the complexity parameters -see fig. (3)- that locates, at the end of the studied period, all the three complexity estimators close to 10 if we apply the random sentence generator to Peter's corpora, and close to 9 in the case of Carl's one. Furthermore, a very interesting feature is found at the first stages of Peter's randomized version of production: The graph is not generally bipartite -specially in the third corpora, whose average chromatic number is about 4 in the model- but so it is in the real case. This is the footprint of the two-words stage grammar which, even displaying an important generative increase, imposes severe constraints on what is actually plausible in a syntactic structure. But this feature is also explicit in latter random versions of the corpora, where the strong difference between all three complexity estimators -in some cases, the chromatic number of the randomized ensemble is twice the real one -see fig (2)- tells us that the complexity of the compatibility rules underlying syntactic relations has an important impact on global combinatorial patterns.

However, we should be aware that the null model is constrained by statistical invariants which are not topological, for the topology of the net is derivative from the statistics over productions. Therefore, the ensemble of networks obtained from the null model could display important divergences from the obtained from real data and, thus, having an implicit impact on the parameters we are studying. Previous works have shown that the topological divergences are weak, but it is known that the studied parameters strongly depend on the connectivity of the nets. To better understand the nature of the observed deviations, we analyzed the behavior of the chromatic numbers against the mean connectivity -see fig. (4). Again, we observe that the chromatic numbers of the networks obtained from the null model of string generation display higher values than the ones we obtained from real data, even in the case where the connectivity is close. This effect is not as strong as the divergences on the chromatic number, but reinforces the idea that the chromatic number is capturing essential combinatorial properties of the underlying system. Furthermore, both the comparison between the size of the giant connected component and the average degree show much more uniform behavior around a quasi linear regime in the case of the randomized ensembles, whereas in the real nets the values are not so clearly distributed.

IV. DISCUSSION

In this study we explored the evolution of the chromatic number in the successive syntactic networks obtained during the acquisition process. The intrinsic combinatorial nature of syntactic relations and the powerful indicator of the internal constraints provided by the

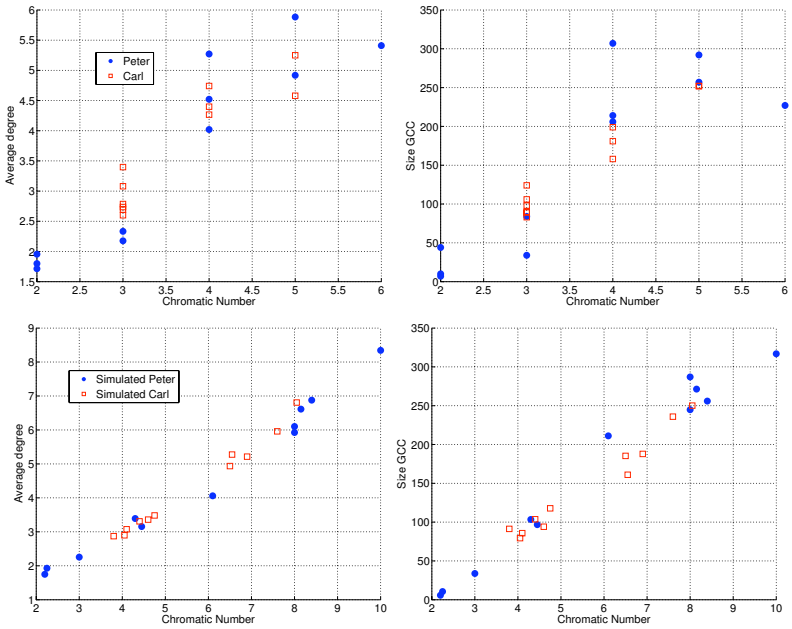


FIG. 4 Relation of the chromatic number and the size and mean connectivity of the nets. The above plots show an interesting, non-trivial deviation of the real data against the obtainde by running the random sentence generator. Indeed, the increasing trend of real chromatic numbers against size and connectivity displays a relation of about $2/3$ if we look at the trend observed for the network obtained by running the null model. This implies that the constraints imposed by grammar affects all scales, being the reduction of the chromatic number about $2/3$ from the randomized one. The above plots confirm that the divergence of the chromatic numbers among real and randomized versions of the system is relevant. It is particularly relevant the clear linear dependence between χ and $\langle k \rangle$ in the left down corner plot.

chromatic number leads us to an interesting picture: in one hand, the chromatic number significantly increases during the acquisition process, a footprint of the increase of the generative power of the underlying grammar and of the emergence of more and more complex rules of sentence construction. On the other hand, the underlying grammar enabling the emergence of complexity also constraints the chromatic features of the net, a feature clearly observed when comparing real nets with the ones obtained from the random string generator. This clearly quantifies how grammar constrains the emergence of links, but a word of caution must be said: the chromatic number establishes communities of non-interacting nodes considering node-node relations, for it is clear that graph representation cannot grasp the hierarchical nature of the syntactic objects. And it is well known that syntactic relations are structure-dependent, not sequence dependent, which means that the obtained results are derivative of a kind of constrains acting at the level of syntactic structure, which is more than a string of dependences, the way that the graph representation forces to use. This observation does not invalidate the obtained results and it can be expanded to any graph representation of language structures, since the graph is the way by which we have a global picture of the global linguistic performance and includes a combinatorial element, which, although being less richer than the syntactic one, provides us relevant information over the system.

But there is another, broader implication of our work. Indeed, as opposed to standard views of community structure the chromatic number defines the minimal structure of communities we can define by considering that no connections among the elements of the same community exist. In these networks, if some kind of compatibility relation is at work in the wiring process, the community structure based on richer connectivity patterns between the members can be misleading, precisely by the fact that, in the real system, elements of the same class cannot be connected. The case of syntactic graphs is paradigmatic, but the partition provided by the chromatic number could shed light to the behavior of many other systems, as a powerful complement to the standard methods of modularity or community structure identification. The statistical significance of such partitions is provided by the sequence of minimal violations we proposed -see eq. (13)- which enables us to evaluate whether, even in the case where we did not find the actual chromatic number, we defined a satisfactory partition in terms of wiring independence. Further works would explore the chromatic number as a good estima-

tor of communities defined by its internal non-interaction as well as the intriguing linear dependencies of the chromatic number with the size and connectivity obtained in the null models in this paper.

Acknowledgments

This work was supported by the James McDonnell Foundation (BCM, SV and RVS) and the Santa Fe Institute (RVS). We thank Complex System Lab members for fruitful conversations.

References

Achlioptas, D., and M. Molloy, 1999, The Electronic Journal of Combinatorics **6**, R29.

Bianconi, G., and M. Marsili, 2006, EPL (Europhysics Letters) **74**(4), 740.

Bickerton, D., 1990, *Language and Species* (University of Chicago Press. Chicago).

Bloom, L., L. Hood, and P. Lightbown, 1974, Cognitive Psychology (6), 380.

Bloom, L., P. Lightbown, and L. Hood, 1975, Monographs of the society for Research in Child Development. Serial 160 (40).

Bollobás, B., 1988, comb **8**(1), 49.

Bollobas, B., 1998, *Modern Graph Theory* (Springer), corrected edition.

Bollobás, B., 2001, *Random Graphs* (Cambridge University Press).

Brooks, R. l., and W. T. Tutte, 1941, Proceedings of the Cambridge Phylosophical Society **39**(2), 194.

Christiansen, M. H., and S. Kirby, 2003, Trends in Cognitive Sciences **7**(7), 300.

Corominas-Murtra, B., 2007, arXiv.org:0704.3708 .

Corominas-Murtra, B., S. Valverde, and R. Solé, 2009, Advances in Complex Systems (ACS) **12**(03), 371.

Ferrer-i-Cancho, R., R. Köhler, and R. V. Solé, 2004, Phys. Rev. E **69**, 051915.

Hauser, M. D., N. Chomsky, and T. W. Fitch, 2002, Science **298**, 1569.

Hristea, F., and M. Popescu, 2003, *Dependency Grammar Annotator, Building Awareness in Language Technology* (Editura Universitatii din Bucaresti).

Ke, J., 2007, arXiv:cs/0701135 .

Maynard-Smith, J., and E. Szathmàry, 1997, *The Major Transitions in Evolution* (University of New York Press. New York).

Radford, A., 1990, *Syntactic Theory and the Acquisition of English Syntax: the nature of early child grammars of English* (Oxford. Blackwell).

Solé, R. V., B. Corominas-Murtra, S. Valverde, and L. Steels, 2010, Complexity **15**(6), 20, ISSN 1099-0526.

Wu, F. Y., 1982, Rev. Mod. Phys. **54**(1), 235.

Zdeborová, L., and F. Krzakala, 2007, Phys. Rev. E **76**(3), 031131.

Corominas-Murtra B. Network statistics on early english syntax: structural criteria. arxiV.org. Cornell University Library. Consultat: 13 set. 2011. Disponible a: arXiv:0704.3708v2

Network statistics on early English Syntax: Structural criteria

Bernat Corominas-Murtra[1]

[1] ICREA-Complex Systems Lab, Universitat Pompeu Fabra, Dr. Aiguader 80, 08003 Barcelona, Spain

Abstract

This paper includes a reflection on the role of networks in the study of English language acquisition, as well as a collection of practical criteria to annotate free-speech corpora from children utterances. At the theoretical level, the main claim of this paper is that syntactic networks should be interpreted as the outcome of the use of the syntactic machinery. Thus, the intrinsic features of such machinery are not accessible directly from (known) network properties. Rather, what one can see are the global patterns of its use and, thus, a global view of the power and organization of the underlying grammar. Taking a look into more practical issues, the paper examines how to build a net from the projection of syntactic relations. Recall that, as opposed to adult grammars, early-child language has not a well-defined concept of structure. To overcome such difficulty, we develop a set of systematic criteria assuming constituency hierarchy and a grammar based on lexico-thematic relations. At the end, what we obtain is a well defined corpora annotation that enables us i) to perform statistics on the size of structures and ii) to build a network from syntactic relations over which we can perform the standard measures of complexity. We also provide a detailed example.[1]. Keywords: Syntax, complex networks, learning, Computation

[1] *This paper is the experimental design of a more extensive work* **The ontogeny of syntax networks through Language Acquisition**, *Corominas-Murtra, B., Valverde, S. and Solé, R. V.*

Contents

1 Introduction **3**
1.1 Different abstractions, different questions: Syntax and Statistical Physics . . . 3
1.2 Aims . . . 4

2 Syntactic networks **4**
2.1 From syntax to networks: what we win and what is lost . . . 4
2.2 Syntactic Networks . . . 5
2.2.1 From syntactic relations to links . . . 5

3 Data **8**

4 Building the Networks of Syntactic Acquisition: Criteria **9**
4.1 Non accepted productions . . . 10
4.2 Accepted Productions . . . 10
4.2.1 Phrases and missing arguments . . . 11
4.2.2 *To be* verb . . . 11
4.2.3 Infra-specification and semantic extension of lexical items 13
4.2.4 First functional particles . . . 14
4.2.5 Duplication of functional words . . . 16
4.2.6 Non-structural lexical items . . . 17
4.2.7 Negation Structures . . . 18

5 Corpora Annotation **19**

6 The average size of structures, $\langle \mathcal{S} \rangle$ **19**

7 Building the Network **20**
7.1 Measures . . . 20

8 Example **22**
8.1 The source . . . 22
8.2 Selected Productions and Analysis . . . 23
8.2.1 XML Format to be read by DGAanotator . . . 24
8.2.2 Selection of *valid* strings, annotation and computation of $\mathcal{S}$ 25

9 Acknowledgments **26**

1 Introduction

In this pages there is an attempt to design and describe a naturalistic experiment on syntax acquisition. Specifically, we want to build a *Syntactic network* in order to study syntax with modern methods of complex network theory. The process is nor standard neither straightforward and deserves to be well described.
There are interesting descriptive frameworks based on networks to study syntax. One of them is the so-called *Dependency grammar*[1]. There are, also, theoretical approaches using graphs. A remarkable member is the *word grammar*[2]. The approach assumed here is closer to the Word-Grammar, despite we develop our own criteria, as well as we consider the graph representation as a linear projection of the constituency hierarchy.
The paper is organized as follows: We firstly discuss the scope and validity of the conceptualization of syntactic relations within a network. The core of the work is devoted to the discussion of the (descriptive) structural criteria to tackle the problem of annotation in early grammars. Finally, a brief compendium of network measures is shown, as well as an illustrating example. All analysis are performed over the PETER corpora of CHILDES database [3] using the DGA-Annotator [4].

1.1 Different abstractions, different questions: Syntax and Statistical Physics

Every abstraction of a natural object implies a particular conception of it in order to answer a specific question. Assuming that every abstraction implies a simplification, we have to explore, then, how different approaches can be complementary or whether some of these approaches are more fruitful than others -i.e., what are the core questions leading to the understanding of such phenomena. Focusing on language, research on syntax seeks to find the minimal set of rules that could generate all -and only- the potentially infinite set of sentences of a given language. Thus, the question addressed by syntax is the problem of decidability or computability of the set of possible sentences of a given language. When dealing with language as a complex network, we have to note that statistical physics works from different perspectives: What are the global features of the dynamics of our system? How the combinatorial space is filled? What is -if any- the role of constraints?
Thus, we don't address questions concerning the structure of the inhabitants -sentences- of our system, but its global dynamics and organization. Note that the questions are different than in the case of syntax: thus, the abstraction we are working in is also different. Note, also, that we are not negating nor denying the particular features of sentence construction. Simply, we work at other level of abstraction. We are confident that information from this different level of approach should be enlightening to questions addressed on grammar itself.
If one wants to apply statistics on some syntactic phenomena, a word of caution is needed because there is a gap between the syntactic procedure and the statistical physics procedure: The former is focused on explaining almost *every*

subtlety of sentence construction, while the latter works on averages over the largest possible set of data. Thus, a compromise has to be assumed because it is not possible to deal with every syntactic phenomena but, also, the statistics has to be built on certain criteria.

1.2 Aims

Thus, the aim of this document is to present a set of descriptive criteria to identify *structure* in early child grammars. This is not a theoretical reflection about the concept *structure* or its evolution during the process of language acquisition. With these criteria, we want to build up the so-called syntactic networks from early child grammars. Indeed, even though many features of the language acquisition process have been identified and well studied, there is a lack of a clear concept of what it is structured or not in early child grammars, namely, there is not a concept such as grammaticality [5] or *convergence* [6], defined in adult grammars. If we take the adult-grammar concept of grammaticality, we will surely reject almost all of children's productions. But it will not be true that many of these rejected utterances are unstructured at all.
In order to overcome these limitations, we developed a set of descriptive criteria to extract the syntactic network of different sets of the child's utterances belonging to successive time stages of the language acquisition process. As we discussed above, we present these criteria employed in the construction of the associated networks[2].

2 Syntactic networks

2.1 From syntax to networks: what we win and what is lost

Formally speaking, a given language, $\mathcal{L}$ is composed of an arbitrary large, but finite set of lexical items $\mathcal{W}$ - or alphabet, in technical words- and of a restricted set of rules and axioms, Γ. These rules describe how the elements from $\mathcal{W}$ can be combined in order to *1.* obtain sentences of $\mathcal{L}$ or to *2.* decide if a given sequence of elements from $\mathcal{W}$ is a sentence of $\mathcal{L}$ or not [7]. Syntax properties are, thus, indicators of grammatical complexity[3]. If one intends to develop a syntax theory to decide whether a given sequence of words -namely, Russian words- is a sentence of *Russian* language, one needs to develop rules involving hierarchical and long range relations. Moreover, the set of rules must be *generative*, in the sense that they should involve some recursive condition to grasp the potential infinity of sentences generated by Russian grammar [9].

[2]Note that many properties of the networks make sense asymptotically, i.e., many utterances need to be analyzed such that the results acquire statistically significance.

[3]In fact, if we would be able to design the minimal program to describe our system, its size (in bits) would be an index of complexity. See [8]

Thus, we can say, without any loose of generalization, that syntax works at the *local level* of language[4], i.e.it operates at the sentence level, no matter how long the sentence is. Now, we wonder about the global profiles of syntactic relations. Note that the question we want to address is not finding the specific rules needed to generate the possible sentences of $\mathcal{L}$, but we want to take a look at the system as a whole. This could seem bizarre when considered from the point of view of mainstream theories of syntax, but it is a common procedure in statistical physics. Global profiles can provide information about general dynamics and constraints acting over the whole system as a complex entity. The unexpected profile given by Zipf's law is an example of global behavior of language dynamics [10].
A note must be added concerning the naturalistic character of this kind of experiments. Syntax has been related with competence abilities. But statistical and naturalistic works are carried out over performance data. Thus, we are inferring the global patterns of performance by assuming some competence abilities.

2.2 Syntactic Networks

Networks revealed as an interesting abstraction to explore the global behavior and dynamics of complex real systems made from units and the associated relations between such units. Let's explore such abstraction for syntax relations. A network $\mathcal{G}(V, E)$ is defined by the nodes V and the links E relating the nodes V [11]. These links can be directed or undirected; we will use the directed ones, if the contrary is not indicated. To build a syntactic network, the mapping of $\mathcal{L}$ onto a graph will be straightforward for the set $V \rightarrow \mathcal{W}$ i.e., the set of lexical items of $\mathcal{L}$ will be the set of nodes of $\mathcal{G}$. The mapping from Γ to E is not so obvious and needs further considerations.

2.2.1 From syntactic relations to links

As we discussed above, the syntactic rules needed to generate any natural language revealed considerable degree of complexity. Thus, it is clear that the statistical treatment employed here is an approximation. Modern syntax is based on recursive operations of *merge* and *move* [6]. Such operations lead the syntactic derivations to display hierarchies and long range relations. These are features that cannot be captured explicitly by a descriptive framework based on linear relations among lexical items -a network approach. But we are approaching the language structure from the point of view of statistical physics: we want to capture the global patterns of the system, thus we cannot specify *all* the local properties. This is contrary to the procedure employed in the Ising models of ferromagnetism, despite the success of this approach is universally acknowledged. Thus we have to decide what is the most essential structure in a

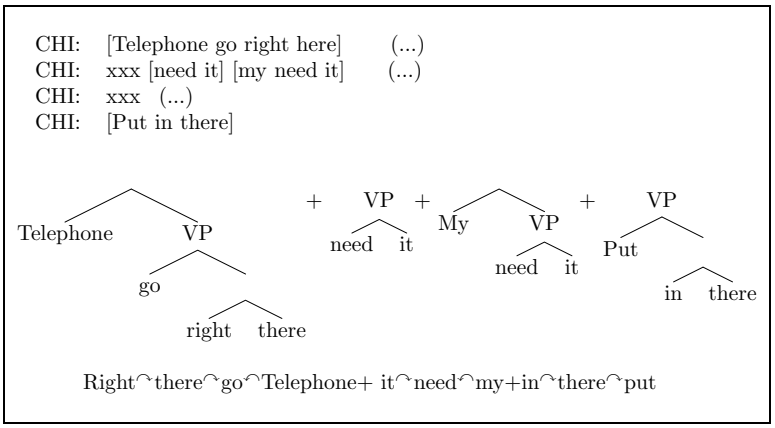
[4]We are not considering the usual *locality* of syntactic relations as understood in many works of syntax, we use the term *local* to specify that syntax operates at the level of individual elements of a given language $\mathcal{L}$

syntactic derivation. We assume that the most fundamental thing one can say from the syntactic point of view about a sentence is its *constituent structure*. Constituent structure can be captured by linear relations. In the following, we define an exact mapping from a hierarchical binary tree to a graph[5], an entity made of binary relations (see figure (1)):

1. Find the basic syntactic structure of constituents without labels nor internal operations, with clear distinctions of complements and the head in every phrase. Detect the verbs in finite forms.
2. Trace an arc from the complement to the head of the phrase. If the complement of a given phrase is also a phrase, trace and arc from the head of the internal phrase to the head of the external phrase. We want to recover the merging order.
3. The head will be the semantically most relevant item.
4. The verbs in finite forms are the head of the sentence.

With the above criteria, we make an attempt to manage data with the less aggressive criteria. Moreover, these assumptions don't constrain us to one or other linguistic school and grasp reasonably with the observed syntactic development of children.
With this method, we do not restrict our set of sentences to the one generated by finite combinatorics. We allow our sentence to be arbitrary long. Thus, our model is only finite because real data is finite in nature, but it doesn't negate the theoretical possibility of infinite generativity, a property expected for any approach of syntax [12]. Moreover, the relevance of the statistical physics properties generally is found in systems asymptotically large.

[5]In the approach of Word-Grammar, the projection of hierarchical structures into linear dependencies is just the inverse of what we adopted here. But it is, essentially, the same procedure. For more information, see [2]



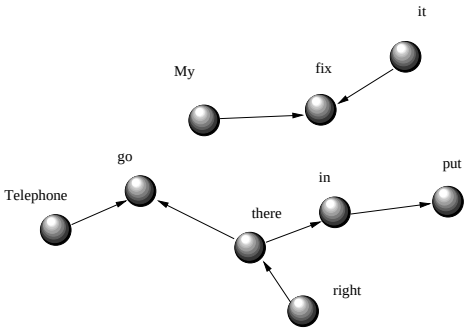


Figure 1: *Building syntactic nets from children free speech corpora.* **A***) We have the transcript of a conversation and we select only child's productions. We identify the structured strings . The notion of structure and the used criteria is widely developed in further considerations.* **B***) Basic analysis of constituent structure, identifying the verb in finite form (if any) in different phrases.* **C***)Projection of the constituent structures into lexical dependencies (note that the operation is reversible: We can rebuild the tree from the dependency relations.).* **D***)Following the dependency relations found by projecting the naked syntactic structure we build, finally, the graph.*

Syntactic networks can be built by other procedures. Dependency syntax [1] has been used in other works [13]. In such an approximation, syntactic networks have been built up by assuming syntactic relations as dependency relations among lexical items. Dependency grammar generates a graph to describe the sentence structure and it is the reason why it is interesting to build networks.

The mechanism to build large nets is straightforward[6].

3 Data

Studies on language acquisition can be divided into two main types: *experimental* and *naturalistic*. The experimental ones are focused on child's response to well-established situations in order to obtain data of some specific trait. Naturalistic studies, at the other hand, are based on child's free speech corpora. These corpora can be extracted, for example, from a recorded session where the child speaks with adults spontaneously. [15]. Our study is clearly naturalistic, and this label takes here its whole meaning, because the procedures to build up *biological* networks, for example, are conceptually, the same.
Data has been extracted from the well-known *CHILDES Database*[7] [3, 16, 17]. The chosen corpus is the *Peter* Corpus, from Bloom 1970. We choose this data for many reasons: 1) Time intervals are regular (about 2 or 3 weeks). 2) Extension of the corpora can be considered large enough to seize global properties, taking into account the intrinsic small size of the system. There is a little exception in corpus 2, which is, by far, the smallest one. Fortunately, this corpus does not seem to belong to a key stage in grammar evolution. 3) The acquisition stages of Peter seem to be the standard ones observed in language acquisition. Thus, it is reasonable to think that our results will not be biased to strange deviations of the particular case study.
Working data includes the 11th first corpora of Peter's 20 corpora. The age period goes from 1 year and 9 months to 2 year and 4 months. As we said above, the aim of the study is to observe whether and to what extent syntactic networks can provide information on the process of language acquisition. The so-called *syntactic spurt* ([15]), which appears later than the *lexical spurt*, is clearly observable in the chosen corpora. Thus, we manage data that begins when the lexical spurt has already taken place and ends when syntactic structures of child's productions are complex enough to be compared with the adult ones. This does not exclude the possibility of more abrupt changes in more advanced acquisition stages, but we stop our analysis here.

Material contains several conversations between adults and the child (These adults are, mainly, researchers and Peter's parents). We selected the child's productions and we studied them considering the discursive context where such

[6]some authors assume the network abstraction for syntax as ontological, i.e., not as an approximation to a complex system of rules involving recursive structures and non-terminal nodes(see [13], [14]). This is not the view adopted here: The network in our approach only is an attempt to grasp some evolutionary features of the system, properties that can be captured by taking a global view to the system, something that is difficult to achieve when looking at the local structure of syntactic relations. Here, networks do not substitute the decision/computation rules because some key features of the syntax itself, such as constituent hierarchy or movement, cannot be treated properly by the graph theoretic abstraction

[7]http://talkbank.org

Corpus	**Age**	**Corpus**	**Age**
1	1;9.7	7	2;1.0
2	1;9.21	8	2;1.21
3	1;10.15	9	2;2.14
4	1;11.7	10	2;3.0
5	1;11.21	11	2;3.21
6	2;0.7		

Table 1: Age of Peter in successive corpora (**years;months.days**). Data from childes database [3, 16, 17]

utterances have been produced. This enables us to *clean* the data. What it means is that we will discard 1) imitations from adults 2) non-structured utterances. A complete explanation of the criteria to accept productions and by this implying that they contribute to the syntactic graph is reported in the next section *Criteria*.
A final note concerning the data: it seems clear that the morphological nature of English, with poor inflectional features makes the identification of functional items easier than in a language with richer inflectional features. The global impact of the morphological nature of a given language on network topology cannot be denied [18], but the global reorganization process observed in child syntactic networks seems to go beyond these singularities[8].

4 Building the Networks of Syntactic Acquisition: Criteria

We selected the productions that allow us to identify some syntactic structure. Obviously, the word *criteria* is due to the evidence that despite the fact that most of early child-productions are not grammatical in the sense of full convergence or complete feature checking, it is not true that they have no structure. Thus, the work of the linguist consists in identifying the clues of syntactic structure in child's productions. Selection is not easy at all, as there does not exist an explicit definition of syntactic structure in early grammars. We considered that there exists structure if there exists, at least, some lexico-thematic relation between the elements in a production. This is the basis of syntactic structure of early English grammars [15]. More complex relations, involving functional words, appear later and syntactic structure can be more easily identified. This is coherent with the observed nature of early grammars.

[8]obviously, words as fundamental units is an intuitive but rather arbitrary choice. Thus, the same study could be extended by considering morphemes as the fundamental unit. This is, maybe a more reasonable choice. In this way, it could be possible to detect more similarities when comparing the acquisition processes of different languages.

4.1 Non accepted productions

First of all, we discarded some transcribed strings if: 1)they are simply an onomatopoeia with no structural role (in some cases *choo choo* could replace *train*). 2)they are non transcribed items -because we supposed it was not possible to understand what the child said. We choose not to consider any of these unidentified lexical items (transcribed in the corpora as *xxx* or *yyy*) in order to ensure the transparency of data managing.

These non-accepted elements are: *a* (in some specific contexts), *ah, an* (in some specific contexts), *awoh, ay, hey, hmm, huh, ka, ma, mm, mmhm, oh, oop, oops, ow, s* (in some specific contexts) *sh, ssh. ta* (in some specific contexts) *uh, uhhuh, uhoh, um, whoops, woo, yum.* **Onomatopoeia**: *choo, Moo, Woof, Bee Bee*

The case concerning *a*, the schwa, will receive a particular attention below.
Some onomatopoeia appear together with its corresponding lexical item. To analyze them, we assume onomatopoeia to be nonexistent. Take, for example:

Peter 9*I want ta write the choo choo train → I want ta write the train*

Considerations related to other non-trivial interpretations, such as the role of *ta*, are extensively developed in the following lines.
In addition, and following the enumeration of non-accepted productions, we find the general case where no structure is identified in a production. In this situation, we consider the utterance as a string of isolated lexical items. Consequently, no links but only nodes corresponding to the lexical items are added to the graph.
More attention has to be paid to imitations. The reason to consider imitations as unacceptable productions is that we have no confidence that such string is identified as a structured one or, simply, as a single lexical element. Imitations are identified by analyzing the discursive context. Some utterances of surprising complexity for its corresponding stage are produced after an *untranscribed adult conversation*: we cautiously removed from the graph such contributions. In Peter 5 corpus, *I can't see it* is produced after an adult conversation and it is, by far, the most complex production of this corpus. It strongly suggests that this is an imitation from something said in such untranscribed adult conversation.

4.2 Accepted Productions

As we stated above, structured productions and lexical items are taken into account. Now we state another assumption: If in the whole utterance we cannot

find global structure but there are some structured strings, then we take these structured strings separately, (see figure 1).

4.2.1 Phrases and missing arguments

In the pre-functional stage (identified in our data until corpus Peter 6) there appear a lot of utterances where only thematic relations seem to be considered by the syntactic system of the child. Thematic relations are fundamental at the syntactic level, and their appearance indicates presence of sub-categorization mechanisms in child grammar. No traces of more complex structure -like agreement- is found in this early stage of acquisition. We consider as syntactic relations the thematic relations between verb and arguments. Moreover, subject elision is usual, due mainly to the facts that 1)utterances are in imperative mode or 2)there is no fixation yet of parametric variation associated to the explicit presence of subject in English. Productions of this kind are:

Peter 5 *Open box*, instead of *Open the box*, (the determiner is missing.)

Peter 5 *wheel walk* instead of *The wheel walks*, (3-singular English agreement is missing)

Peter 6 *two truck* instead of *two trucks*, (no plural agreement)

This leads to the logical conclusion that productions like **open the* will not be accepted. The reason is clear: if we assume thematic relations as the basic building blocks of child syntax, the non-presence of the semantically required argument but its determiner is not enough to define any relation.
Relations between verbal head and functional words are specifically considered in phrasal verbs. Its isolated production is considered a structured utterance. Several reasons support our choice: 1) Their intrinsic complex nature, 2) We cannot conclude that there are lexicalized imitations because, in adult speech, phrasal verbs usually are *broken* by a noun or determiner phrase:

Turn [the wheel]$_{SD}$ *out.*

4.2.2 *To be* verb

Semantically vacuous predications (those which involve the *to be* verb) are often produced without realization the verb. We argued that missing arguments or lack of agreement in a production could be not the only reasons to conclude that there is not any structure in child utterances. This was justified because strong semantically items were present in discussed productions. The case of copulative constructions will be treated close to the ones involving missing functional words. In this case, no presence of the verb does not motivate the consideration

of non-structured production. An interesting production is:

Peter 5 *Wheels mine* Instead of **The** *wheels* **are** *mine*

In this case we have a predication *mine* from something *Wheels*. Formally, *are* is a semantic link between predication and the element from which something is predicated [**?**]. So the missing of the *to be* verb could be considered analogous to the missing of a functional particle. The same situation arises from:

Peter 7 *That my pen*

Usually, when inflectional morphology appears, some infinite forms are present without the finite form of the *to be* verb. This is the case of some present continuous utterances such as:

Peter 8 *I writting*

This case should be treated as the above case: There is some predication with semantic structure. Just the opposite is also found: presence of the *to be* verb with an infinitive or finite form:

Peter 8 *I'm write too*

In this case, we could assume that the child is acquiring inflectional morphology and that this utterance is a present continuous one without inflection. In the other hand, we could consider that *'m* has not a role in the sentence and thus, this can be treated as a single finite sentence *I write too*.
Analogously,

Peter 7 *I'm do it*
or
Peter 6 *cars goes away*

Are treated as single finite sentences: *I do it* and *The cars go away*
Some lexicalized phrases in adult language, such as *back seat* or *thank you*, have been considered as a complex structures. The reason is to be coherent: If we assume that *fix it* is clearly an imperative structured sentence, at this stages of acquisition there is no reason to think that *back seat* or *thank you* have to be considered differently. Moreover, this interpretation is also coherent with the one developed for phrasal verbs.
A special case of imitations involving the *to be* verb will be accepted. These imitations involve some *adaptation* of adult syntax to the syntax in which the child is competent. An example should be:

(Peter 6)

Adult: *Is that a truck?*

Child(1): *That's a truck?*

Child(2): *That a truck?*

In this example, adult production involve an interrogative sentence with subject inversion. The first imitation (*Child*(1)) retains all the lexical items but the sentence is translated as interrogative without subject inversion. In the second successive imitation (*Child*(2)) the verb is missing. But the elements to define a predication are still at work -with a *schwa* as a determiner, suggesting that the child is entering into the functional stage.

4.2.3 Infra-specification and semantic extension of lexical items

During the acquisition process, extension of meaning is subject to variations. To know which is the intrinsic nature of these changes is not our aim, but we have to manage such situations. Thus, we find utterances where the child uses in the *wrong* way some lexical item that could be related semantically with the *right* lexical item. As an example:

Peter 5 *More screwdriver*

Which could, checking the context, be properly replaced by constituents or lexical items with related meanings:

Another screwdriver

or
screw it again

or
screw it more (or *harder...*)

In the first case, we could consider that the child made some semantic extension of the word *more* and it has enough traces to define a syntactic relation. But context can lead us to a second or third interpretation. Generally, if there is a great ambiguity we reject such utterances as structured ones. In this case, we should not consider any syntactic structure. Thus, we don't define any relation in productions such as:

Peter 5 *Screwdriver help*

Peter 5 *More [fix it]*$_{SV}$ (We don't define any relation between *More* and the SD *fix it*)

The semantics of the productions are intuitive, but is hard to justify clearly some kind of syntactic dependency.
Strings displaying mistakes in the use of personal pronouns and possessives have considered as structured. Generally, we could associate such mistakes to the absence or weakness of case system. But many productions, as we reported above, have structure without any trace of case assignation. Examples of this kind of utterances are:

Peter 5 *My fix it* instead of *I fix it*

Peter 8 *Me write* instead of *I write*

This assumption is reinforced by realizing that, in some cases, a production with wrong pronoun is repeated correctly without any conversational pause:

Peter 8 *Me found it (...) I find it*

This situation cannot be confused with the missing of the *to be* verb such in the case of *wheels mine*. This case has to be considered as above mentioned when dealing with missing *to be* verb structures.

4.2.4 First functional particles

In early corpora (1-4) child productions display very poor structures. This is the so-called *pre-functional* stage, where no functional words appear in structured productions. Beyond this point, some lexical elements -we are mainly talking about the *a*, the *schwa*- seem to act as a *proto*functional particles. Whether this schwa has a phonological or functional-syntactic character is an open question [19, 20].
Some authors related the presence of these items as one step to combinatorial speech [20], but they realized that, in early stages, the role of these items is more related to phonological processes of language acquisition, without any functional or structural role, at the syntactic level. Other authors such as Veneziano & Sinclair "*linked these phenomena more specifically to the child's development of grammatical morphemes considering them as a sort of an intermediate form on the way to grammatical morphemes.*" [19]pp 463. Roughly speaking, we can say that the core of this reasoning is rooted in the idea that the role of such items is dynamic, going at very first stages as *filler syllabes* without any syntactic role and acquiring grammatical features during the process to end as functional particles, with specific syntactic role.
The lack of consensus around a topic that seems to be crucial in syntactic acquisition theorizations forces us to be really cautious in interpreting such items. Furthermore, functional words such *a* are strongly candidates to be the hubs in a fully developed syntactic network. Hub are the most connected nodes on a network, being, thus core pieces in network organization. Every candidate

belonging to the set of functional particles is specially analyzed in order to discard simple phonological phenomena. Thus, for every occurrence of such items there will be an individual decision, taking in account the context and with the framework defined by Veneziano & Sinclair [9].
Specifically, we considered that sometimes the *schwa* plays a functional role. It is reasonable to assume, thus, that sometimes the *schwa* is substituting a specific functional particle. In this cases we assume that the *schwa* acts within the syntactic structure as the substituted particle. Several examples can illustrate such reasoning:

Peter 6 *Light* **a** *hall*

Peter 6 *light in* **a** *hall*

Peter 6 *look* **a** *people*

In this case, it seems that *a* substitutes *the*. **a** should be treated as a determiner. This is a very difficult choice, because purely phonological interpretation could be enough to justify the presence, specially in the third case.
Sometimes choice is really ambiguous. Take for example:

Peter 6 *There* **a** *new one*

Such a case **a** could be easily interpreted as a pure phonological phenomena. But if we consider the vacuous semantic nature of the *to be* verb, we could understand these occurrences as protofunctionals. We removed these most ambiguous cases. We also rejected as unstructured utterances productions involving confuse sequences of functional particles as:

Peter 6 *Will an a in there*

Any interpretation is really confusing.
There are cases where the presence of the **a** is clearly purely phonological. For example:

Peter 6 *more get* **a** *more*

Peter 6 a *ride a horsie*

Peter 5a *this thumb*

Peter 7 *hmmm my a*

[9] In the Veneziano & Sinclair's study, the chosen language is French, but we take as general some conclusions that seem to coincide with the observed phenomena in English acquisition

Beyond the non-definition of personal pronouns due to the weakness of the case system, we find the pronoun *I* as an *a*:

Peter 7 *a want milk*

Peter 7 *a want ta get out*

An interesting sequence of that reinforces our considerations is:

Peter 7 *a put it on (...) my put it on*

Finally, it is interesting to note the presence of elements that are, to some extent, a mixing between *a* and *to*: *ta*. The occurrence of this particle is rare and located explicitly at the very beginning of the funcional stage. The remarkable fact lies on the evidence that is located where it should be the preposition *to*. This could imply that in fact there is a transition from a pure phonological role to a functional one.

$a \rightarrow ta \rightarrow to$

Thus, we interpret *ta* as an intermediate stage but, due to its location within the sentence and the context, we assume it behaves as a preposition:

Peter 6 *[Have [ta [screw it]]$_{PP}$]*

Peter 7 *[Have[ta [screw it]]$_{PP}$]*

The emergence of English syntax is strongly tied to the emergence of functional particles. This is the reason why we decided to take into account this kind of lexical components: despite almost every utterance involving such items can be object of many considerations, there are enough motivation to try to define a descriptive criteria to deal with them.

4.2.5 Duplication of functional words

It is usual to find, at the beginning of the functional stage, that a verb that subcategorizes, for example, a prepositional phrase, display two successive prepositions:

Peter 6 *Look* **at in** *there*

To manage this kind of productions we assumed, first, that these imply that the child conceives[10] a syntactic structure that involves prepositional phrases.

[10] *Conceives* implies that the child is competent in this ind of productions, thus we are not using this verb in terms of explicit knowledge

Following this reasoning, we make the following structural description:

Look **at in** *there* → *[Look [at there]$_{PP}$]*

In that case, *in* is interpreted as an independent lexical item. Not only prepositions are involved in this duplication phenomena, but also determiners:

Peter 9 *One that screwdriver*

Interpretation rules out *one* as a member of any structure, leading the SD *[that screwdriver]$_{SD}$* alone.
A situation analogous to famous one described by Braine (p.160-161) [22] is:

Peter 7 *Get* **another** *one paper* → *Get another paper*

Thus, as above, determiner duplication is not considered in the structural analysis.

4.2.6 Non-structural lexical items

By this name, we designate the lexical items that are present in a conversational framework but cannot be explicitly interpreted as members of some syntactic structure, such as *Hello*, or *Ok*. The reason to include these elements as connected to the network is due mainly because their are produced in non-arbitrary context. Thus, we assume that they linked to the first element of the sentence they precede:

Peter 5 *Ok Patsy*

Obviously, previous reasons are at work when dealing with such items. Thus, the conversational context has to be analyzed to interpret these items. For example, In the following situation, *bye* has not been considered as a member of any structure. The reason is that it is produced among analogous expressions, leading it to interpret more in a pragmatic sense:

Peter 7 *see you, bye, see you*

Sequences of numbers or other elements produced as a list are not considered as members of any structure:

Peter 7 *one two three...*

Sometimes, personal nouns are produced by the child to demand attention from adult people. In these situations, we do not accept them as a members of structured sentences.

Some residual cases to be commented are the ones related with strings of nouns:

Peter 9 *Piece tape...*

Which are clearly unstructured, if conversational context does not conspire in the other way. Sequences like

Peter 9 *off on tv...*

are ruled out as structured ones because any structure proposal leads us to a certainly bizarre sentence in terms of meaning and because there is a lack of many elements that can act as clues to find some structure. Thus, they are considered as isolated lexical items:

Peter 7 An *Jenny*

4.2.7 Negation Structures

When the functional stage is being consolidated, we find more complex structures. Among others, interrogatives involving subject inversion or negation structures.
Negation structures sometimes imply the presence of the auxiliary *to do* are produced using the negative particle alone:

Peter 7 No *put it here*

This context suggests us that *No* could be replacing the auxiliary form *don't*. *Don't put it here* is its grammatical counterpart. Nevertheless, we consider *no* as replacing *don't* and, thus, as a member of a bigger syntactically structures utterance. Obviously, as we said above, context has to rule out interpretations such as **No**, *put it here*. Analogous structures can be:

Peter 7 No *ride a bike*

There are other situations where we considered suitable not to consider *no* as a member of any structure:

Peter 8 *in the bag* **no**

We cannot conclude that there is a syntactic relation among the negation operator and some other lexical item of the string. Maybe a parametric [21] hypotheses could save this production by suggesting that the location of the negation operator within the structure may be a parametric feature. Despite interesting, we choose the rule these productions out for reliability purposes.

Furthermore, at the same time, there are productions like:

Peter 9 No *in this box*

suggesting that the child *knows* the ordering of negation structures in English. In the latter case, as before, we considered not risky to identify *no* as a member of a bigger structures sentence.

5 Corpora Annotation

Previous set of criteria enables us:

1. To identify and characterize syntactic structures in early child language.
2. To project them into word-word dependencies in order to annotate the corpus.

The corpora is annotated *by hand*. This enables us to be accurate and to manage ambiguous situations. The program used to perform the annotation is the so-called **Dependency Grammar Annotator** (DGA annotator). This program was developed by Marius Popescu [4] from the University of Bucaresti and has a nice and easy interface. It works with XML files, whose internal structure will be described in the example of the last section.

6 The average size of structures, $\langle \mathcal{S} \rangle$

With these criteria in hand, we are ready to perform a first analysis of grammar complexity. Such an analysis is closed to the classical MLU[11]. What we can compute, now is the average size of syntactic structures. Thus, in a production, we can, for example, find two syntactically unrelated structuress_1 and $s_2 s$. The number of lexical items of these structures will be its sizes $|s_1|$ and $|s_2|$. Such an utterance will contribute to the computation of $\langle \mathcal{S} \rangle$ with two structures.
For example, in

Look **at in** *that*

we have two structures:
s_1 =[Look, [at, that]] $\rightarrow$ $|s_1| = 3$
s_2 =in $\rightarrow$ $|s_2| = 1$
$\langle s \rangle = (3+1)/2 = 2$
(Note that single words are considered as size-1 structures) To obtain $\langle \mathcal{S} \rangle$ The average is computed over all utterances. Such a measure will provide us clues to decide whether the size of productions has information about grammatical complexity. Its evolution can be related with working memory limitations.

[11] Medium lenght of utterances, often measured on utterance size in words or morphemes

7 Building the Network

Once we analyzed a conversation, we can build the network. The process is as follows. Due to the nature of our analysis, we will have a collection of words, which define the set $\mathcal{W}$:

$$\mathcal{W} = \{car, it, ...\} = \{w_1, w_2, ..w_n\} \tag{1}$$

This define the set of nodes of our network. If, during the conversation, we find some structure where two words w_i, w_k are related syntactically -using the above criteria!- we say that $w_i rightarrow w_k$[12] and that there is a link $w_i \to w_k$.

$$\mathcal{E} = \{car \to want, it \to want, it \to fix...\} = \\ \{w_1 \to w_k, w_2 \to w_k, ...\} \tag{2}$$

Remark that:
All the words and links only appear once a time. This enables us to separate -as far as possible- some contextual deviations from the specific conversations. Also, there can be many isolated nodes.
Finally, we compute the adjacency matrix $\mathcal{A}_{ij}$. This matrix is the representation of the graph and the abstract object where all computations of graph complexity are performed. If the child produced n different words during the conversation, the size of this matrix will be, obviously n^2.
The adjacency matrix of the directed graph will be:

$$\mathcal{A}_{ij} = \begin{cases} 1 \leftrightarrow w_i \to w_i \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

If we consider the undirected version of this graph, $\mathcal{A}^u_{ij}$ will be defined as:

$$\mathcal{A}^u_{ij} = \begin{cases} 1 \leftrightarrow w_i \to w_j or w_j \to w_i \\ 0 \text{ otherwise} \end{cases} \tag{4}$$

Note that $\mathcal{A}^u_{ij}$ is symmetrical, whereas $\mathcal{A}_{ij}$ it is not.
Now we are ready to perform an exhaustive analysis of network complexity.

7.1 Measures

A first and fundamental question we find when dealing with such measures is whether the net is made of a large number or small, isolated graphs or if it displays a clearly differentiated Giant Connected Component (GCC) that contains most of the connected words -i.e. words syntactically active in some production. The number of words contained on such a component or its relative size are interesting statistical indicators. Strikingly, from the very beginning, child's syntactic graphs display a clear and very differentiated GCC. For mathematical purposes, we will use the matrix representation of the connectivity pattern of

[12] Do not confuse it with the logical conditional

the GCC, the so-called *adjacency matrix*. An element of such a matrix is $a_{jk} = 1$ if there exists a link among the words W_j and W_k and $a_{jk} = 0$ otherwise. If the contrary is not indicated, -we will compute the following measures over the GCC of our graphs.
The number of links (or *degree*) $k_i = k(W_i)$ of a given word $W_i \in \mathcal{W}$ gives a measure of the number of (syntactic) relations existing between a word and its neighbors. The simplest global measure that can be defined on Ω is the average degree $\langle k \rangle$. For the T-th corpus, it will be defined as

$$\langle k \rangle_T = \frac{1}{N_w(T)} \sum_{W_i \in \mathcal{W}} k(W_i) \tag{5}$$

where $N_w(T)$ indicates the number of words present in the T-th corpus. This number is known to increase through acquisition in a steady manner. This and other measures are computed on the largest component of the graph.
Beyond the average degree, two basic measures can be used to characterize the graph structure of the GCC of the T-th corpus. These are the average path length (L_T) and the clustering coefficient (C_T). The first is defined as $L_T = \langle D_{min}(i,j) \rangle$ over all pairs $W_i, W_j \in \mathcal{W}$, where $D_{min}(i,j)$ indicates the length of the shortest path between two nodes. Roughly speaking, a short path length means that it is easy to reach a given word $W_i \in \mathcal{W}$ starting from another arbitrary word $W_j \in \mathcal{W}$. The second is defined as the probability that two vertices (e.g. words) that are neighbors of a given vertex are neighbors of each other. In order to compute the clustering, we define for each word W_i a neighborhood Γ_i. Each word $W_j \in \Gamma_i$ has been syntactically linked (via the above defined projection) at least once with W_i in some sentence. The words in Γ_i can also be linked among them, and it is what the clustering coefficient evaluates. The clustering $C(\Gamma_i)$ of this set is defined as

$$C(\Gamma_i) = \frac{1}{k_i(k_i - 1)} \sum_j \sum_{k \in \Gamma_i} a_{jk} \tag{6}$$

and the average clustering of the GCC concerning the T-th corpus is simply $C_T = \langle C(\Gamma_i) \rangle$. The clustering C provides a measure of the likelihood of having triangles in the graph. Concerning the average path length, for random graphs with Poissonian structure we have

$$D = 1 + \frac{\log [N/z_1]}{\log [z_2/z_1]} \tag{7}$$

being z_n the average number of neighbors at distance n. For Poissonian graphs, where $z_1 = \langle k \rangle$ and $z_2 = \langle k \rangle^2$, we have the following approximation: $D \approx \log n / \log \langle k \rangle >$. It is said that a network is a *small-world* when $D \approx D_{random}$ (and clearly $D \ll N$). The key difference between a Poissonian network and a real network is often $C \gg C_{random}$ [23].
Another quantity of interest is the degree of affinity among nodes with the same connectivity. In this way, the behavior of hubs is specially relevant, as well as

they organize the overall structure of the net. A network is said to be *assortative* if hubs tend to be connected among them. At the other side, a network is said to be *dissassortative* if hubs tend to avoid connections among them. Language networks at different scales display a high degree of dissassortativeness [13]. To quantify the degree of assortativeness, we use the so-called Pearson's coefficient for nets [24]:

$$\rho = \frac{c\sum_i j_i k_i - \left(c\sum_i \frac{1}{2}(j_i + k_i)\right)^2}{c\sum_i \frac{1}{2}(j_i^2 + k_i^2) - \left(c\sum_i \frac{1}{2}(j_i + k_i)\right)^2} \tag{8}$$

where j_i and k_i are the degrees of the edges at the ends of the ith edge with $i = 1, ..., m$, $c = \frac{1}{m}$ and being m the number of edges. If $\rho < 0$ the net is dissassortative, whereas if $\rho > 0$ the net is assortative.

8 Example

Below we have a fragment of the conversation transcribed in the Corpus *Peter 7*. We will detail the analysis that we perform. Firstly, we show the source corpus. We follow by selecting Peter's productions. After that we select the structures and analyze this structures and we tag them. We finish by computing $\langle \mathcal{S} \rangle$ of this fraction of text and by showing the obtained net.

8.1 The source

*PAT: hey Pete that's a nice new telephone looks like it must do
everything it must ring and talk and .
%mor: co—hey n:prop—Pete pro:dem—that v—be & 3S det—a adj—nice adj—new
n—telephone
n—look-PL v—like pro—it v:aux—must v—do pro:indef—everything pro—it
v:aux—must
v—ring conj:coo—and n—talk conj:coo—and .
%exp: Peter has a new toy telephone on table next to him
%com: ¡bef¿ untranscribed adult conversation
*CHI: xxx telephone go right there .
%mor: unk—xxx n—telephone v—go adv—right adv:loc—there .
%act: ¡bef¿ reaches out to lift phone receiver, pointing to place where
wire should connect receiver and telephone
*MOT: the wire .
%mor: det—the n—wire .
*PAT: oh ¡the & te¿ [//] the wire's gone ?
%mor: co—oh det—the n—wire v:aux—be & 3S v—go & PERF ?
%com: ¡aft¿ untranscribed adult conversation
*CHI: xxx need it my need it xxx .
%mor: unk—xxx v—need pro—it pro:poss:det—my n—need pro—it unk—xxx
.

%act: ¡aft¿ goes to his room on Mother's suggestion, returns with wire
*CHI: xxx .
%mor: unk—xxx .
*PAT: uhhuh .
%mor: co—uhhuh .
*LOI: why don't you bring your telephone down here Peter ?
%mor: adv:wh—why v:aux—do neg—not pro—you v—bring pro:poss:det—your n—telephone
adv—down adv:loc—here n:prop—Peter ?
*LOI: why don't you put it on the floor ?
%mor: adv:wh—why v:aux—do neg—not pro—you v—put & ZERO pro—it prep—on det—the n—floor ?
%act: ¡aft¿ Peter puts it on floor ¡aft¿ Peter is trying to attack "wire"
to phone and receiver
%com: ¡aft¿ untranscribed adult conversation
*LOI: what're you doing ?
%mor: pro:wh—what v—be & PRES pro—you part—do-PROG ?
*CHI: 0 .
%act: ¡aft¿ Peter goes to hall closet, tries to open it
*MOT: what do you need ?
%mor: pro:wh—what v—do pro—you v—need ?
*CHI: xxx .
%mor: unk—xxx .
(...)
*CHI: put in there .
%mor: v—put & ZERO prep—in adv:loc—there .
%act: attaching wire to phone
*LOI: ok it's all fixed oops it was out all fixed there .
%mor: co—ok pro—it v—be &3S qn—all part—fix-PERF co—oops pro—it v—be & PAST & 13S
adv—out qn—all v—fix-PAST adv:loc—there .

8.2 Selected Productions and Analysis

To work with the DGA Annotator, we need, firstly, to extract the child's productions. To do this, we programmed a routine in PERL language able to extract child's productions. Below there is a simple pseudocode as a sample:

```
FILE=PETER_k
for(i=5; i<=LONGFILE; i++)
{
   if(FILE[i]= /PETER/)
   {
   j=j+1;
   PETER[j]= "FILE[i]";
```

```
   }
}

for(i=0; i<=LONGFILE; i++)
{
      @PETER[i]= tr/*PETER:/ /;
      @PETER[i]=  tr/./ /;
      @PETER[i]=  tr/,/ /;
      @PETER[i]=  tr/;/ /;
      @PETER[i]=  tr/:/ /;
      @PETER[i]=  tr/!/ /;
      @PETER[i]=  tr/</ /;
      @PETER[i]=  tr/>/ /;
      @PETER[i]=  tr/?/ /;
      @PETER[i]=  tr/Â¿/ /;
      @PETER[i]=  tr/*/ /;
}
```

If we apply the above algorithm to the sample of text of the example, we obtain:

xxx telephone go right there
xxx need it my need it xxx
xxx
0
xxx
put in there

8.2.1 XML Format to be read by DGAanotator

Further we need to provide the obtained strings of words with a XML format, in order to manage them with the DGA Annotator. Below we have an example of the string *put in there*.

```
⟨?xml version="1.0" encoding="iso-8859-1"⟩
⟨!DOCTYPE DGAdoc SYSTEM "dga.dtd"⟩
⟨DGAdoc⟩
⟨s⟩
   ⟨tok⟩
       ⟨orth⟩put⟨/orth⟩
       ⟨ordno⟩1⟨/ordno⟩
   ⟨/tok⟩
   ⟨tok⟩
       ⟨orth⟩in⟨/orth⟩
       ⟨ordno⟩2⟨/ordno⟩
   ⟨/tok⟩
   ⟨tok⟩
```

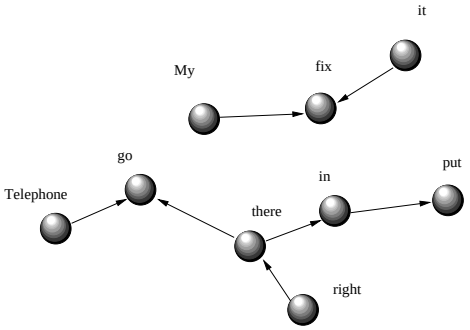


Figure 2: Graph of the sample.

```
        ⟨orth⟩there⟨/orth⟩
        ⟨ordno⟩3⟨/ordno⟩
    ⟨/tok⟩
⟨/s⟩
⟨/DGAdoc⟩
```

8.2.2 Selection of *valid* strings, annotation and computation of $\mathcal{S}$

We reject *xxx* and 0 as lexical items and proceed to annotate with the DGA anotator.

Once the corpus is annotated (with the criteria developed through the paper!) we generate the set of words. This can be done by sampling the XML file once annotated by using a routine close to the ones shown above (PERL or Python are the ideal languages). To compute graph parameters and more mathematical artifacts, it is a good choice to use a *stronger* language, such as C or C++.

$$\mathcal{W} = \{\text{telephone, go; right, there, need, it, my, put, in}\} \tag{9}$$

And the analysis is, roughly speaking:

$$s_1 = [\text{telephone}[\text{go}[\text{right there}]_{\text{PP}}]_{\text{VP}}]_{\text{TP}} \quad |s_1| = 4 \tag{10}$$

$$s_2 = [\text{need it}]_{\text{VP}} \quad |s_2| = 2 \tag{11}$$

$$s_3 = [\text{my}[\text{need it}]_{\text{VP}}]_{\text{TP}} \quad s_3 = 3 \tag{12}$$

$$s_4 = [\text{put}[\text{in there}]_{\text{PP}}]_{\text{VP}} \quad s_4 = 3 \tag{13}$$

Thus, we can compute $\langle \mathcal{S} \rangle$:

$$\langle \mathcal{S} \rangle = \frac{4+2+3+3}{4} = 3 \tag{14}$$

and, following the criteria developed above, we can define $\mathcal{E}$

$$\begin{aligned} \mathcal{E} = \{&\text{telephone} \rightarrow \text{go},\ \text{right} \rightarrow \text{here},\ \text{here} \rightarrow \text{go}; \\ &\text{it} \rightarrow \text{need},\ \text{my} \rightarrow \text{need},\ \text{there} \rightarrow \text{in},\ \text{in} \rightarrow \text{put}\} \end{aligned} \tag{15}$$

We have built the graph.

9 Acknowledgments

My indebtedness for Carlos Rodríguez-Caso for his ideas and his patient advising in teaching PERL. I also want to acknowledge Harold Fellermann for his help in programming python routines, Andreea Munteanu for the careful reading of the manuscript and Barbara Soriano for its useful comments on language acquisition process. My indebtness, also, to Ricard V. Solé and Sergi Valverde as well as this work is a part of a bigger work where their ideas have been illuminating. Finally, I'd like to acknowledge Ramon Ferrer i Cancho for his contributions at the beginning of the work.

References

[1] Melçuck (1989) *Dependency Syntax: Theory and Practice.* University of New York, New York

[2] Hudson, R. (2006). Language networks. The new word grammar. New York: Oxford University Press.

[3] MacWhinney, B. (2000). The CHILDES project: Tools for analyzing talk. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

[4] Popescu, M. (2003) *Dependency Grammar Annotator*, Building Awareness in Language Technology (editors: Florentina Hristea & Marius Popescu). Editura Universităţii din Bucureşti, 2003, p.17-34.

[5] Chomsky, N. (1965) *Aspects of the theory of syntax.* MIT Press. Cambridge, MA.

[6] Chomsky, N. (1995) *The minimalist program.* MIT Press. Cambridge, MA.

[7] Chomsky N. (1963) Formal properties of grammars. In *Handbook of Mathematical Psychology* Vol II. Wiley Publishing Company, New York

[8] Chaitin, G. J. (1987) *Algorithmic information Theory.* Cambridge University Press. Cambridge, MA.

[9] Chomsky N. (1963) Formal properties of natural languages. In *Handbook of Mathematical Psychology* Vol II. Wiley Publishing Company, New York

[10] Zipf, G. K. (1949) *Human Behaviour and the Principle of Least-Effort* Addison-Wesley. Cambridge MA.

[11] Bollobàs, B. (1998) *Modern Graph Theory.* Springer. New York.

[12] Chomsky, N. & Miller, G. (1963) Finitary models of language users. In *Handbook of Mathematical Psychology* Vol II. Wiley Publishing Company, New York

[13] Ferrrer i Cancho, R., Solé, R. V. & Kohler, R. (2004) Patterns in syntactic dependency networks. Phys. Rev. E. 70 056135.

[14] Ferrer i Cancho, R. (2005) The structure of syntactic dependency networks: insights from recent advances in network theory. In *The problems of quantitative linguistics*, Altmann, G., Levickij, V. & Perebyinis, V. (eds.). Chernivtsi: Ruta. pp. 60-75.

[15] Radford, A. (1990) *Syntactic Theory and the Acquisition of English Syntax: the nature of early child grammars of English* Blackwell, Oxford.

[16] Bloom, L.; Hood, L. & Lightbown, P. (1974) Imitation in language development: If, when and why. Cognitive Psychology, 6, 380-420.

[17] Bloom, L.; Hood, L. & Hood, L. (1975) Structure and variation in child Language. Monographs of the society for Research in Child Development, 40, (serial No. 160)

[18] Sole, R. V., Corominas Murtra, B. & Valverde, S. *Unpublished Work*

[19] Veneziano, E. & Sinclair, H. (2000) The challenging status of filler syllabes on the way to grammatical morphemes.

[20] Bloom, L. (1970) *Language development: form and function in emerging grammars* (Research Monograph No. 59) The MIT Press, Journal of Child Language, 27, 461-500.

[21] Hyams, N. (1997) *Language Acquisition and the theory of Parameters.* Dordrecht: Reidel

[22] Braine, M. D. S. On two types of models of the internalization of grammars. In D. I.. Slobin (ed.), *The ontogenesis of grammar: a theoretical symposium.* Academic Press (New York)(1971) pp. 153-186

[23] Dorogovtsev, S. N. & Mendes, J. F. F. (2003) *Evolution of Networks: From Biological Nets to the Internet and WWW*, Oxford University Press, New York.

[24] Newman, M. E. J.(2002) Assortative mixing in Networks. Phys. Rev. Lett. 89, 208701

Chapter 6

PAPERS ON THEORETICAL SYNTAX

Fortuny J, Corominas-Murtra B. Some formal considerations on the generation of hierarchically structured expressions. CatL. 2009; 8: 99-111.

The backbone of theoretical syntax

Bernat Corominas Murtra[1,2]
[1]ICREA-Complex Systems Lab, Universitat Pompeu Fabra (GRIB), Dr Aiguader 80, 08003 Barcelona, Spain
[2]Institut de Biologia Evolutiva CSIC-UPF. Passeig martim de la Barceloneta

We explore the set-theoretical basis of syntactic relations and conclude that current syntax theory can be consistently organized and supported by set theory. Specifically, we demonstrate that merge operation can be identified with set union operation and we conclude that syntactic objects are *nests*, the core concept of order theory. Going further, we analyze the unavoidable finite-size effects demonstrating that the emergence of cyclic movement is a natural consequence, and needs not longer be postulated. Formal features are mathematically introduced, enriching the structural hallmark provided by merge. The final outcome is the so-called *nesting grammar*, a set-theoretical construct that has the aim of being richer and wider enough to properly act as the backbone of theoretical syntax. The generative power of the obtained grammars is studied, obtaining preliminar results related to the location of nesting grammars -according to their properties- within the Chomsky hierarchy. We observe that the nested structures offer a rigorous conceptual alternative of what has been intuitively referred as "recursion" in linguistics literature. Finally, we emphasize that the mathematical construct is intended to be minimal.

Keywords: Syntax, Order Theory, Merge Operation

I. INTRODUCTION

The emergence of language is one of the great transitions in evolution (Smith and Szathmáry, 1995). The identification of *what happened* to the small group of hominids leading to this astonishingly powerful communication system is hard to identify. What seems clear is that at some point we humans or a close precursor- began to be able to construct in an apparent effortless way arbitrarily deep hierarchical structures to carry meaning and, thereby, to be used as a communication system(Chomsky, 2008; Hauser *et al.*, 2002; Smith and Szathmáry, 1995). There are several ways to attack the problem, but the most urgent business is to define *what is the problem*, for it is clear that human language is an object resulting from the complex interaction of many subsystems, each of them having their internal dynamics and particular properties. Recent developments in theoretical syntax isolated an operation that might be at the core of this biological innovation: The operation *merge* (Chomsky, 1995; Radford, 1997; Rosselló, 2006). This operation refers to a mental ability that concatenates two sets to generate a bigger one which, in turn, could be concatenated and so on. The evolution of neural activity presumably reached a point where this operation was possible or, more correctly, as argued in (Hauser *et al.*, 2002), the conceptual system and the articulatory system where developed enough to predate this computational ability for communication.

Under the framework proposed above, the study of core syntax is the study of how this operation acts as a generative mechanism, how it is constrained, or how it constrains the set of possible linguistic objects, all of this under *reality* conditions. This operation would generate structures to be compositionally interpreted by the conceptual system. Under this view, syntactic theory would set merge operation as the central concept from which the theory is built. The objective of this article is to develop a mathematical framework putting *merge* both at the starting point and at the center of the theory. This defines a formal architecture based on set theory which is intended to be the formal basis underlying syntactic theory.

Indeed, throughout the paper we will defend that the mathematical basis underlying syntactic theory is set theory. Within this mathematical hallmark, merge operation is just set union

$$"\bigcup"$$

between two sets (Fortuny and Corominas-Murtra, 2009). The successive application of merge defines an order, and it is worth to note that this issue is more or less formally stated in (Chomsky, 2008). However, the rigorization of merge operation provided in (Fortuny and Corominas-Murtra, 2009) demonstrated that the set-theoretical interpretation of the merge operation generates a special kind of sets common to order theory, namely *nests* , and that such ordered sets where the fundamental object to describe syntactic concepts and relations.

Given a set $W = \{w_1, ..., w_n\}$, a nest of W is a collection of subsets of W, $W_1, ..., W_k \subset W$ such that can be ordered by inclusion (Kelley, 1955; Kuratowski, 1921; Suppes, 1972):

$$W_1 \subset W_2 \subset ... \subset W_k.$$

The aim of this work is to go one step beyond from (Fortuny and Corominas-Murtra, 2009), thereby advancing in the exploration of the possibilities of a mathematically consistent theory of core syntax. We first set up the hallmark proposed in (Fortuny and Corominas-Murtra, 2009) with some -non fundamental- variations. Within this framework, it is shown that nests are the kind of objects generated by *merge* operation and that they grasp

all structural properties intuitively attributed to a given syntactic object. The generation of such objects is performed through an abstract machine to be informally referred as *nesting machine*. The finite nature of the machine must be assumed in force, and this results in a memory limitation which, when studied, has as a consequence the emergence of cyclic movement. We then explore how to equip, in a rigorous and consistent way, the generated grammar with a structure of features, which at the same point, constrains and enriches the kind of emerging grammar. This enables us to generate a family of complex grammars to act as the theoretical support of syntactic theory. Finally, we explore how the features of such grammars determine the expressive power in terms of the Chomsky hierarchy and we how human syntax can be embedded within this framework.

II. THE GENERATION OF STRUCTURES: MERGE

The framework provided here is intended to be mathematically consistent and is a proposal for the backbone of syntax theory. It can be understood as a proposal *radically* minimalist, in chomskyan terms, for it reduces the core syntax relations to relations among sets. The starting point of our proposal is the rigorous definition of the operation *merge* postulated in the minimalist program as the core operation governing the generation of recursive structures. Specifically, all notions of syntax are reduced to relations among a kind of sets called *nests*, which properly describe the structure of a syntactic object. The definition and discussion provided in this section are purely structural and no other notion than *generation of structure* is considered. Special attention is paid to the consequences of the finite nature of the *machine* generating such structures. This section presents a revised form of a previous paper written jointly with Jordi Fortuny (Fortuny and Corominas-Murtra, 2009).

A. The identification of nests and syntactic objects

Given an alphabet, viewed as a finite set of singletons,

$$A = \{\{a\}, \{b\}, ..., \{z\}\},$$

we define a machine that works as follows: At the first step, s_0, this machine generates the set M_0, which is an element of A. At the following step, s_1, it generates a new set M_1, by forming the union of M_0 and a given member of A. At step s_n we generate the set M_n, which is the union of M_{n-1} and an element of A. When an arbitrary element of A, namely, $\{k\}$, comes into the computation at step s_i, its element, k, becomes an occurrence, k_i, thereby distinguishable from other occurrences of the element $\{k\}$ during the derivation process. This can be summarized through a recursive schema:

$$\begin{aligned} M_0 &= \{a_0\} \\ M_n &= \{k_n\} \bigcup M_{n-1}, \end{aligned}$$

where n is unboundedly large, but not infinite.The record of these merging operations generates the following set N:

$$N = \{M_0, ..., M_n\}.$$

N is a *nest*, i.e., a family of sets linearly ordered by the inclusion relation (Kelley, 1955; Kuratowski, 1921; Suppes, 1972):

$$M_0 \subset M_1 \subset M_2 \subset ... \subset M_n$$

Now consider the more general case where we want to allow the nesting machine to perform $X \bigcup Y$, when X is a set M_j generated by the nesting machine at step s_j and Y a singleton whose element is not a primitive element of A but already a non-trivial nest generated in parallel. To enable such operations, we need to postulate that several machines are working in parallel. We shall label the different machines at work as $D_1, D_2, ..., D_n$. The idea is that several nests can be produced in parallel and act as singletons generating other nests. For the sake of clarity, the outcome of the union operation between two sets built by the j-th nesting machine at step s_k will be labeled as M_k^j. When $\{a\}$ has been merged at the step s_k by the nesting machine D_j it becomes the occurrence $\{a_k^j\}$. At some point, D_i generates N_i and stops, which means that N_i cannot grow anymore, feeding another active machine. At the last step of a given computation, only one machine D_j can be at work. This implies that all nesting machines that worked in the computation generated their outcomes as inputs for other nesting machines. We observe that this condition is the realization of the so-called *single root condition* for syntactic structures (Partee *et al.*, 1990). It is important to remark that the step label "$_k$" is not dependent on the particular machine where the specific computation is performed, i.e., it is an universal "clock". Therefore, a given nest generated at D_i can have, as starting element $\{a_k^i\}$, with $k > 0$.

The definition of the *nesting machine* shows how structures are generated by only considering the role of the merge operation, interpreted as the union operation between two sets. The (infinite) set of all nested structures generated by the nesting machine is $\mathcal{F}_{\mathcal{L}}^*$.

B. Set relations and Syntactic relations

The objects generated by the nesting machine have an internal structure where several relations among their building blocks can be identified. We observe that the set-theoretical nature of the framework we are developing forces these relations to be defined in terms of set relations. As we shall see, core concepts of syntax theory like *constituency* or *dominance* (Radford, 1997; Rosselló, 2008) can be naturally defined in terms of inclusion to a given set.

1. Constituency

Now we shall concern ourselves with the problem of giving a consistent and rigorous definition of *syntactic constituent*, and therefore construct the *set of syntactic constituents* forming a given structure by which the single root condition holds -i.e., where at the last stage of the computation only one machine remains open and all other machines produced an output which fed another nesting machines. This is the core concept of the developed theory.

Let N_i be the nest which is the whole outcome of the nesting machine D_i, created during $s_j, ..., s_k$ steps, i.e.,

$$N_i = \{M_j^i, M_{j+1}^i ..., M_k^i\},$$

having

$$\mathbf{larg}(N_i) = M_k^i$$

as the largest element[1],. Then, we build the set C_i, namely:

$$C_i = \{c_0^i, c_1^i, ..., c_k^i\},$$

where c_i^j's are all the nests obtained if we stop the generation of N_i at a given step of the computation. Formally, $c_j^i = \{M_0^i, M_1^i, ..., M_j^i\}$ $(j \leq k)$. It is clear that:

$$\mathbf{larg}(c_j^i) = M_j^i,$$

furthermore, C_i also has an element which is the largest one, c_k^i. We finally observe that, consistently:

$$c_k^i = N_i.$$

There is another class of constituents of a given nest N_i generated by the nesting machine D_i, after k steps, namely, all the elements that, at any step j have been merged to M_{j-1}^i. This is the set of all individual elements of the derivation, i.e., $\bigcup N_i$. We observe, however, that, since N_i is a nest,

$$\bigcup N_i = \mathbf{larg}(N_i)$$

which is precisely M_k^i. Therefore, the elements of C_i and $\mathbf{larg}(N_i)$ actually define the *set of constituents of* N_i, to be named $\mathcal{C}_i$. Putting it formally,

$$\mathcal{C}_i = C_i \bigcup \mathbf{larg}(N_i)$$

Therefore, the *set of all constituents of* a given N, where we take into account all the derivational record potentially involving $D_1, ..., D_n$ nesting machines to be written as $\mathcal{C}$, is defined as:

$$\mathcal{C} = \bigcup_{D_1,...,D_n} \mathcal{C}_i.$$

An element $\alpha \in \mathcal{C}$ is a *constituent* of N.

We have rigorously characterized syntactic constituent as a special kind of set directly generated by the application of merge.

Remark Notice that a constituent can be either a nest containing more than one element or a single element of the alphabet. It is NOT a constituent, for example, the set

$$X = \{M_3^i, M_5^i, ..., M_j^i\}.$$

We finally define the sets

$$\tilde{C} = \bigcup_{D_1,...,D_n} \mathbf{larg}(N_i), \quad C = \bigcup_{D_1,...,D_n} C_i,$$

and

$$M = \bigcup_{D_1,...,D_n} N_i.$$

These sets will be useful in the forthcoming developments. Notice that, generally:

$$\tilde{\mathcal{C}} \bigcap \mathcal{C} \neq \varnothing.$$

2. Dominance

Dominance is here defined as a binary relation between the set of non terminal nodes M and the constituents, $\mathcal{C}$. To define dominance in a given nest N, we need to construct an auxiliary set relation accounting for immediate dominance relations, R, such that:

$$R \subset M \times \left[M \bigcup \tilde{\mathcal{C}}\right],$$

and defined as: $(\forall \alpha \in \left[M \bigcup \tilde{\mathcal{C}}\right])$,

$$R(M_i^k, \alpha) \leftrightarrow (\alpha \in \{M_{i-1}^k, M_i^k \setminus M_{i-1}^k\}). \qquad (1)$$

Relation R refers to the relation of immediate dominance between two abstract sets ordered by inclusion as the result of the successive operation of merge. Therefore, the *domain* of a given $M_i^k \in M$ will be properly defined from the transitive closure of R, to be referred as[2] $T(R)$. The domain relation is defined as an order relation[3] Δ between non-terminal nodes and constituents, namely:

$$\Delta \subset M \times \mathcal{C}.$$

---

[1] Given a set A, the *largest* element of A $\mathbf{larg}(A)$ -if any- is the set α by which $(\forall \beta \neq \alpha \in A)(\beta \subset \alpha)$. In a nest, the existence of the largest element is holds directly by definition (Kelley, 1955).

[2] Given a set relation $F \subseteq A \times B$, the transitive closure of F, $T(F) \subseteq A \times B$ is a transitive relation defined as follows:$(\forall \langle a, b\rangle, \langle b, c\rangle \in F) \Rightarrow (\langle a, b\rangle, \langle b, c\rangle, \langle a, c\rangle \in T(F))$.

[3] Throughout this paper, order relations will be those relations $F \subset A \times B$ by which **i)** $(\langle a, b\rangle \in F) \Rightarrow (\langle b, a\rangle \notin F)$ **ii)** $\forall \langle a, b\rangle, \langle b, c\rangle \in F) \Rightarrow (\langle a, b\rangle, \langle b, c\rangle, \langle a, c\rangle \in F)$ **iii)** $\langle a, a\rangle \notin F$.

which is defined as $(\forall \alpha \in \mathcal{C}), (\forall M_i^k \in M)$:

$$\Delta(M_i^k, \alpha) \leftrightarrow (\langle M_i^k, \mathbf{larg}(\alpha)\rangle \in T(R)). \quad (2)$$

Therefore, the domain of a given constituent κ, $\Delta(\kappa)$ is defined as[4]:

$$\Delta(\kappa) = \{\alpha \in \mathcal{C} : \langle \mathbf{larg}(\kappa), \mathbf{larg}(\alpha)\rangle \in T(R)\}. \quad (3)$$

The dominance relation is the natural order relation emerging from the nest structure generated by the successive application of merge operation.

C. Internal Merge

In the above lines we considered that merge operation is actually the set-union operation between nest-like sets. Now we introduce the possibility to *copy* elements already present in the structure. Since syntactic objects are formed by copies -either from the alphabet or from a given derivational space- of nests obtained by merge operation -whose members could also be nests-, it is natural to include the copies of parts of the structure as potential elements to be merged at a given step of the computation. We explicitly introduce the possibility to merge copies of *parts* of the structure to the structure itself, a feature postulated for syntactic structures called *internal merge*, or *movement* (Radford, Chomsky, Hornstein). We explicitly don't want to make any distinction between different types of merge, therefore, our definition of merge will include both internal and external merge as particular, not qualitatively different, cases of merge.

Let us be more formal. Let D_m be the m-th nesting machine acting in parallel in which the operations of merge are being performed, and $\alpha \in \mathcal{C}$ be an already formed constituent in any of the nesting machine that have been at work until step s_k. By *internal merge* we define the operation of copying α, thereby obtaining the instance of α, α_k, to be merged to M_{k-1}^m, and, therefore, obtaining M_k^m:

$$M_k^m = \alpha_k \bigcup M_{k-1}^m.$$

If this constituent is copied again to be merged at step $k_1 > k$, it generates the instance α_{k,k_1}. Therefore, if a constituent is internally merged w times, the different copies will be labeled as:

$$\alpha_k, \alpha_{k,k_1}, ..., \alpha_{k,k_1,...,k_{w-1}}, \alpha_{k,k_1,...,k_{w-1},k_w},$$

being $k < k_1 < ... < k_{w-1} < k_w$.

The successive application of internal merge by copying the same constituent generates a *chain* of copies of the constituent α, written as $CH(\alpha)$ is a set whose elements are sets ordered by inclusion -i.e., a nest- depicting the successive w copies of the constituent that have been internally merged to generate a given syntactic object, namely:

$$CH(\alpha) = \{\{\alpha\}, \{\alpha, \alpha_k\}, ..., , \{\alpha, \alpha_k, ..., \alpha_{k,k_1,...,k_{w-1}}, \alpha_{k,k_1,...,k_{w-1},k_w}\}\}.$$

We highlight two facts: 1) the nest concept is again crucial to rigorously define, in an unified way, the kind of objects we find in syntactic derivations. In the above case, it perfectly encodes the concept of *chain*. 2) If we assume that merge is performed between *already* formed syntactic objects, either the trivial ones that live in the set A or the more complex ones generated in other nesting machines working in parallel, there is no reason to deal with constituents already generated by merge in a different way. Therefore, internal merge is a natural consequence of the definition of merge. We stress that, even counterintuitive, it is not to enable internal merge what introduces artifacts.

III. FINITE MEMORY AND CYCLICITY

In this section we impose the restrictions due to finite memory and we explore the consequences we can derive from them. It is worth to emphasize that to consider a potentially infinite memory storage is completely nonsense in a physical world. Therefore, the restrictions exposed in this chapter have to be assumed *in force*. The existence of syntactic phenomena produced by memory restrictions have been postulated in the past -see (Uriagereka, 2002)-, in particular, within the recent framework of the theory of phases (Chomsky, 2008). Within our framework, this memory restriction impacts in two, different but strongly related, ways. The first natural bound to be defined is the maximum number of nesting machines that can be at work *simultaneously*. Therefore, one has to assume that there exists a constant $\mu \in \mathbb{N}$ restricting the number of active nesting machines:

$$D_1, ..., D_\mu, \quad \mu \in \mathbb{N}. \quad (4)$$

There are many reasons to take it for granted. The first one is related to the intrinsic finite nature of any natural machine performing the merge operation. The second one relates to the internal coherence of the theoretical construct, for it provides a decidibility criteria by which, in a finite number of steps, one can reject a given structure as a member of our language.

Finite memory has an intimate relation with internal merge: As we shall see, memory limitations can act as a trigger for the emergence of such a phenomena. In the previous section (section II.C) internal merge we define the merging at a given state of the computation of

[4] We observe that, given a rigorous definition of dominance, it is straightforward to give a rigorous definition of c-command. Indeed $\alpha \in \mathcal{C}$ c-commands $\beta \in \mathcal{C}$ if

$$(\exists \theta, \gamma \in \mathcal{C}) : (R(\gamma, \alpha) \wedge R(\gamma, \theta) \wedge (\langle \mathbf{larg}(\theta), \mathbf{larg}(\beta)\rangle \in (T(R)))).$$

We do not develop the properties of c-command here.

an already formed constituent. Therefore, the interpreter somehow has to store the whole structural information in the working memory to properly copy a part of the structure and merge it to the top. Since the potential size of the structures is unbounded, not to have any restrictions for the size of constituents that can be internally merged at a given computational step would imply a potentially unbounded working memory, which is non reasonable. Consistently, we have to introduce a critical window of accessibility from the step of the derivation we are in. The most important effect of this memory limitation is the emergence of cyclic movement -postulated within all the current frameworks of generative grammar (Chomsky, 1995; Hornstein, 2000; Radford, 1997; Rosselló, 2008; Uriagereka, 2002)- as the natural solution of an unavoidable problem of memory finiteness. Successive cyclicity would emerge as the natural solution to the problem that arises when the interpretation of some constituent needs to be performed at two positions of the structure, as it happens with *wh*-questions with the object. In our framework, cyclic movement would emerge from the need of the constituent to be always at the *window* defined by the working memory. To properly formalize how it impacts to the structure, we need a bit more notation.

Let us define the function $\sigma : \mathcal{C} \to \mathbb{N}$. This function acts in the following way: given a constituent $\alpha \in \mathcal{C}$, $\sigma(\alpha)$ returns the step in which this constituent α has been generated. Specifically,

$$\sigma(\alpha) = \begin{cases} j \ \text{ iff } \ \alpha = c_j^k \in C \\ j-1 \ \text{ iff } \ \alpha = \mathbf{larg}(c_j^k) \setminus \mathbf{larg}(c_{j-1}^k) \in \tilde{C} \setminus C \end{cases} \tag{5}$$

With this quantity, a definition of distance between constituents naturally arises. Indeed, let N be a nest formed by successive merge involving one or more nesting machines, whose set of constituents is $\mathcal{C}$. Let $\alpha, \beta \in \mathcal{C}$. We define the *distance between two constituents* $\alpha, \beta \in \mathcal{C}$, $\mathbf{d}(\mathcal{C} \times \mathcal{C}) \to \mathbb{N}$, as[5]:

$$\mathbf{d}(\alpha, \beta) = |\sigma(\alpha) - \sigma(\beta)|. \tag{6}$$

Now suppose that we want to internally merge a copy of an already created constituent α, α_k to the constituent β thus creating the new constituent γ. Suppose that this merging has no memory limitations. Since the potential application of merge is unbounded, any constituent created before the step we are considering is *available* or ready to be copied and internally merged. Therefore, we necessarily conclude that the internal memory of the machine is potentially unbounded, for it is able to recover any constituent created at any step of the computation. This is, obviously, nonsensical. Thus, we are forced to assume that there is an upper bound on the distance by which a constituent α can be extracted from the structure, copied and internally merged to β. Or, in other words, there is a window from the last step of the derivation to a certain depth of the structure by which any constituent created *before* is completely unaccessible. In formal terms, $(\exists \delta_c \in \mathbb{N})$ such that:

$$(\gamma = \langle \alpha_k, \beta \rangle) \leftrightarrow (\mathbf{d}(\alpha, \beta) < \delta_c)\,. \tag{7}$$

In this equation, δ_c refer to the upper bound on the distance of constituents that can be internally merged to generate another constituent. In other words, it defines a window of accessibility.

Condition (7) has, as a consequence, the emergence of cyclic movement. We present it as a lemma, since it is an important consequence of the assumptions of the framework.

Lemma 1 (*Emergence of cyclicity*). *Let γ a constituent generated by the union (merge) of constituent β and a copy of constituent α. Assume that condition (7) holds and that $\mathbf{d}(\alpha, \beta) > \delta_c$. Then, there is a chain of intermediate copies $CH(\alpha)$,*

$$\begin{aligned} CH(\alpha) \ = \ & \{\{\alpha\}, \{\alpha, \alpha_k\}, ..., \\ & \{\alpha, \alpha_k, ..., \alpha_{k,k_1,...,k_{w-1}}, \alpha_{k,k_1,...,k_{w-1},k_w}\}\} \end{aligned}$$

such that $w = \sigma(\beta)$ and, if $\alpha_{k,k_1,...,k_\ell}$ and $\alpha_{k,k_1,...,k_{\ell+1}}$ are successive copies:

$$\mathbf{d}(\alpha_{k,k_1,...,k_\ell}, \alpha_{k,k_1,...,k_{\ell+1}}) < \delta_c.$$

Proof. We will proceed in a constructive way. Suppose, as stated in the lemma, that the constituent α is generated at some step s_j of the derivation and that, for interpretative reasons -which are not our competence- a copy of it must be merged to β to generate γ in such a way that

$$\mathbf{d}(\alpha, \beta) > \delta_c.$$

As well as the syntactic object grows, we reach a point in which constituent κ is generated, in such a way that

$$\sigma(\kappa) < \sigma(\beta); \quad \text{but} \quad \mathbf{d}(\alpha, \kappa) = \delta_c - 1.$$

(Notice that, when the syntactic object is fully developed, $\kappa \in \Delta(\beta)$ -see definition of domain and equation (3).). Therefore, if we want α -or a copy of it- to be available in further computational steps, we have to copy it thus generating $\alpha_{\sigma(\kappa)}$. Otherwise, α will be a part of a *frozen* constituent, and there it no longer be possible to extract it from the constituent ρ such that $\mathbf{d}(\rho, \beta) = \delta_c - 1$. We can expand this reasoning until we reach a point whose distance to β is lower than δ_c. This naturally generates a chain of copies

$$\begin{aligned} CH(\alpha) \ = \ & \{\{\alpha\}, \{\alpha, \alpha_{\sigma(\kappa)}\}, ..., \\ & \{\alpha, \alpha_{\sigma(\kappa)}, ..., \alpha_{\sigma(\kappa),k_1,...,k_{w-1}}, \alpha_{\sigma(\kappa),k_1,...,k_{w-1},k_w}\}\} \end{aligned}$$

---

[5] To certify that $\mathbf{d}(\mathcal{C} \times \mathcal{C}) \to \mathbb{N}$ is a *distance* -we have to check if $\mathbf{d}$ holds the so-called *axioms of distance* (Kelley, 1955): $(\forall \alpha, \beta, \gamma \in \mathcal{C})$

A1 $\quad \mathbf{d}(\alpha, \alpha) = 0$

A2 $\quad \mathbf{d}(\alpha, \beta) = \mathbf{d}(\beta, \alpha) > 0$

A3 $\quad \mathbf{d}(\alpha, \beta) + \mathbf{d}(\beta, \gamma) \geq \mathbf{d}(\alpha, \gamma)$.

It is easy to check that $\mathbf{d}$ verifies the above axioms.

such that $w = \sigma(\beta)$ and, if $\alpha_{\sigma(\kappa),k_1,...,k_\ell}$ and $\alpha_{\sigma(\kappa),k_1,...,k_{\ell+1}}$ are successive copies:

$$\mathbf{d}(\alpha_{\sigma(\kappa),k_1,...,k_\ell}, \alpha_{\sigma(\kappa),k_1,...,k_{\ell+1}}) < \delta_c.$$

As stated in the lemma. □

Remark 1. The above lemma describes how the growing of the syntactic object through successive merging operations generates a *sliding window* of availability on what can be internally merged. Successive cyclycity would emerge by the need to have the constituent always available, thereby climbing up the structure as it grows, to be always inside the window defined by δ_c. At this point of the theory, where we are exploring the first consequences of the mathematical backbone of syntax, we are not interested on *why* a given constituent must be copied and merged, this is a problem involving the semantic interpretation. What we demonstrate here is that, if a given constituent must be copied and then merged at an arbitrary step of the computation, then by assuming finite memory -an unavoidable assumption-, successive cyclicity emerges as the natural solution to this tension.

Remark 2. In real languages, successive cyclycity must be constrained by other restrictions imposed by the specific features of the creative process of meaning generation through compositional features. It is sure, thus, that the conditions by which it emerges will be much more constrained and less evident than this simple argument of finite memory. However, this can be postulated as the explanation for the *need* to have this cyclic patterns. Then, how this cycles are performed in a real language, is a matter involving many other factors -specially, compositionality.

We finally observe that, given a finite δ_c, the minimal size of the chain $CH(\alpha)$ is[6]:

$$|CH(\alpha)| \geq \left\lceil \frac{\mathbf{d}(\alpha,\beta)}{\delta_c - 1} \right\rceil.$$

A chain will display minimal size if all two successive copies α', α'' are located exactly at $\mathbf{d}(\alpha', \alpha'') = \delta_c - 1$.

The set generated by successive applications of merge considering finite memory i.e., μ, δ_c finites, is $\mathcal{F}_\mathcal{L}$. Clearly,

$$|\mathcal{F}_\mathcal{L}| = \infty, \text{ and } \mathcal{F}_\mathcal{L} \subset \mathcal{F}^*_\mathcal{L}.$$

IV. MERGE AND FEATURE CHECKING

So far we defined the fundamental operation for the generation of syntactic structures -merge- and we explored the consequences of the finite nature of the machine that generates them -i.e., the consequences of the constraint of finite memory. As we argued, the finite memory constraint is intrinsic and unavoidable and, as we emphasized, not to assume this constraint in addition to the unbounded nature of the merge operation leads us to assume that the system can store an unbounded amount of structural information, which is nonsensical for an object living in the real world. The set of all objects generated by this machinery is $\mathcal{F}_\mathcal{L}$. Now we introduce a new class of constraints, namely *feature checking* (Chomsky, 1995), which are not intrinsic but related to the requirements of the semantic interface to deal with a more restrictive set of structures than the one represented by $\mathcal{F}_\mathcal{L}$. Intuitively, feature checking restricts the operation merge to those cases where some compatibility relation among the sets to be merged is defined. Technically, we equip the elements of the alphabet with an structure of features and a we add to the nesting machine a set of compatibility relations among these features. Therefore, beyond its intrinsic *nest*-like nature, the syntactic objects will have a collection of elements that will define the compatibility relations.

A. Features

Let us define, in an abstract way, the concept of feature[7]. We observe that this concept will not be fully justified until we define the compatibility relations, which is developed in the next subsection.

We will define *feature* as a function

$$\varphi_i : A \to \mathbb{N}. \tag{8}$$

where an abstract property is coded in some way by natural numbers. Once we defined *feature*, the view we had over the elements of the alphabet will change radically. Now the alphabet will be defined by a much more structured set $\mathbf{A}$. Specifically, $\{\mathbf{x}\} \in \mathbf{A}$ will now be defined by an ordered pair between an element of A and a k-tuple (being k finite) of features, i.e.:

$$(\forall \{\mathbf{x}\} \in \mathbf{A}) \quad \{\mathbf{x}\} = \{\langle x, \langle \varphi_1(x), \varphi_2(x), ..., \varphi_k(x)\rangle\rangle\}, \quad \{x\} \in A. \tag{9}$$

Finally, we say that a given $\{\mathbf{x}\} \in \mathbf{A}$ such that

$$(\exists i \leq k) : (\varphi_i(x) = 0)$$

is said to be *neutral* or *stable* with respect the feature φ_i.

---

[6] $(\lceil x \rceil = p \in \mathbb{N}) \Leftrightarrow (p \geq x \wedge p - 1 < x.)$

[7] In this text we reserve the word *feature* to the definition provided in eq. 8. The word feature is used in various ways in linguistic theory. The definition proposed here resembles also the concept of *label* (Chomsky, 2008). However, I hope that the explicit definition provided here avoids any confusion. The word feature is kept in this text without further discussion because actually determines some characteristic of the object. The need to clarify this issue was raised by Jordi Fortuny in a personal communication.

Notice that *features* can be interpreted as the *valence* of a given atom in the periodic table, a numerical information from which we extract the combinatorial properties of such an atom to generate molecules.

1. Compatibility relations

The next task is to detail the role of features within the process of syntactic structure generation. First, we define a function, f, the *checking function*. The *checking function* is a function

$$f : \left(\mathbb{N}\bigcup\{0\}\right)^k \times \left(\mathbb{N}\bigcup\{0\}\right)^k \to \left(\mathbb{N}\bigcup\{0\}\right)^k,$$

by which, $(\forall\{\mathbf{x}\},\{\mathbf{y}\} \in \mathbf{A})$

$$\begin{aligned} f(\mathbf{x},\mathbf{y}) &= f(\langle\varphi_1(x),...,\varphi_k(x)\rangle,\langle\varphi_1(y),...,\varphi_k(y)\rangle) \\ &= \langle f_1(\varphi_1(x),\varphi_1(y)),...,f_k(\varphi_k(x),\varphi_k(y))\rangle \end{aligned}$$

where

$$(\forall i \leq k) f_i : \left(\mathbb{N}\bigcup\{0\}\right) \times \left(\mathbb{N}\bigcup\{0\}\right) \to \mathbb{N}\bigcup\{0\},$$

being defined as

$$f_i(\varphi_i(x),\varphi_i(y)) = \begin{cases} 0 \text{ iff } \langle\varphi_i(x),\varphi_i(y)\rangle \in K_i \\ \varphi_i(x) \text{ otherwise,} \end{cases} \quad (10)$$

where K_i is a set relation, the *set of compatibility relations of feature* φ_i,

$$K_i \subset \mathbb{N} \times \mathbb{N}. \quad (11)$$

Consistently, the set $\{K_1,...,K_k\}$ is the *set of compatibility relations*. Having defined feature and the set of compatibility relations, we are ready to define *structure of features and compatibility relations* which, as we shall see in the forthcoming sections, will be the identity card of a given grammar. Previous to further developments, we observe that, even f_i's describe a behavior of abstract features -or *labels*- close to the one postulated in several works of generative grammar -where the abstract properties are *projected* to an upper level of the structure (Chomsky, 1995)- some degree of arbitrariness can be supposed for eq. (10). Indeed, one could suppose some more complex way to define checking, instead of the simple binary nature of the function proposed. Although it is true that it could be enriched -thereby having a more complicated version of f_i- we consider that this is the simplest form to introduce feature checking. Therefore, a more complex version of it must be exploited only if some unavoidable lack is found. Even in the hypothetical case where it is demonstrated to be insufficient, its simplicity has the pedagogical value that enables us to go further in the theory. We also observe that a change of f changes the way by which feature checking is applied, but the theoretical hallmark is not affected.

The close relation between features and compatibility is unified through the concept of *structure of features and compatibility relations*, Φ, defined as a k-tuple (being k the number of features) of ordered pairs whose first member is the feature and the second one the compatibility relation related to it:

$$\Phi = \langle\langle\varphi_1,K_1\rangle,...,\langle\varphi_k,K_k\rangle\rangle. \quad (12)$$

Remark. The role we attribute to Φ and the way by which it constrains the generation of structures will determine how a given grammar works.

Given two syntactic objects, we can define different *compatibility relations* among their features. Two elements $\{\mathbf{x}\},\{\mathbf{y}\} \in \mathbf{A}$ are *compatible* if

$$(\exists j \leq k) : (\langle\varphi_j(x),\varphi_j(y)\rangle \in K_j).$$

On the contrary, these two elements $\{\mathbf{x}\},\{\mathbf{y}\} \in \mathbf{A}$ are *incompatible* if they are not compatible and

$$(\exists j \leq k) : (\varphi_j(x) \neq 0 \wedge \varphi_j(y) \neq 0).$$

Finally, we say that two elements $\{\mathbf{x}\},\{\mathbf{y}\} \in \mathbf{A}$ are *mutually neutral* if they are not compatible neither incompatible, i.e.,

$$(\forall j \leq k)(\varphi_j(x) = 0 \vee \varphi_j(y) = 0).$$

It is worth to note that such definitions will be naturally expanded to constituents in general, and that they will not only be restricted to relations among alphabet elements.

2. Merge and feature checking

Now we are going to reproduce the merging operation described at the beginning of section II, but now considering the role of features and how they evolve through the process of growing of the structure.

Let us begin in a somehow informal way, in order to grasp the spirit of what we are developing. Suppose that we have a nest generated by the successive applications of merge. At the first step we have

$$M_0 = \{\mathbf{a}_0\}, \text{ being } \{\mathbf{a}_0\} = \{\langle a_0,\langle\varphi_1(a),...,\varphi_k(a)\rangle\rangle\}.$$

Then we generate the constituent whose structural information is summarized in the following nest-like set structure -as usual:

$$c_1 = \{\{a_0\},\{a_0,a_1\}\}$$

but we it is no longer work only with a *structural* constituent, as it has been done in the above sections, but now we must also encode information about features. Therefore, we define the constituent -now $\mathbf{c}_1$- as an ordered pair where the first term encodes the structural

information and the second one is a k-tuple encoding feature information:

$$\mathbf{c}_1 = \langle\{\{a_0\},\{a_0,a_1\}\}, f(\mathbf{a_1},\mathbf{a_0})\rangle.$$

Then we generate the constituent $\mathbf{c}_2$ by the application of merge over $M_1 = \{\mathbf{a}_0, \mathbf{a}_1\}$ and a given element $\{\mathbf{x}_2\} \in \mathbf{A}$. Therefore, we have

$$\mathbf{c}_2 = \langle\{\{a_0\},\{a_0,a_1\},\{a_0,a_1,x_2\}\}, f(\mathbf{x}_2,\mathbf{c}_1)\rangle$$

The crucial step is that, since f has been defined regardless the kind of argument we are dealing -if it is represented by a k-tuple of natural numbers-, it can be written in a recursive way, namely,

$$f(\mathbf{x}_2, c_1) = f(\mathbf{x}_2, f(\mathbf{a}_1,\mathbf{a}_0)).$$

Now we are ready to properly formalize the above intuitions. Let us suppose a nest $\mathbf{N}$ built using a single nesting machine. In presence of an structure of features and compatibility relations Φ, a constituent formed at at step s_k, $\mathbf{c}_k \in C$, is a singleton containing an ordered pair

$$\mathbf{c}_k = \{\langle c_k, \psi(c_k)\rangle\}$$

where c_k is the structural constituent, as defined in section , and $\psi(c_k)$ is the *state* of the constituent, defined as:

$$\psi(c_k) \equiv f(\mathbf{x}_k, \mathbf{c}^i_{k-1}) = f(\mathbf{x}_k, f(\mathbf{y}_{k-1}, (...(f(\mathbf{t}_1,\mathbf{w}_0))...))),$$

where $\mathbf{x}_k, \mathbf{y}_{k-1}, \mathbf{z}_{k-2}, ..., \mathbf{t}_1, \mathbf{w}_0$ are copies of $\mathbf{x}, \mathbf{y}, \mathbf{z}, ..., \mathbf{t}, \mathbf{w} \in \mathbf{A}$. This enables us to can generalize the condition of stability or neutrality of a given constituent: A given constituent $\mathbf{c}_k$,

$$\mathbf{c}_k = \{\langle c_k, \psi(c_k)\}\rangle$$

is said to be *neutral* or *stable* if

$$\psi(c_k) = \langle 0, ..., 0\rangle.\Box$$

Remark 1. $\psi(c_k) = \langle 0, ..., 0\rangle$ is achieved by successive merging operations where (some) elements have compatible features[8].

Remark 2. Let us suppose that $\mathbf{x} \in \mathbf{A}$ is internally merged to c_k, thereby forming $\mathbf{c}_{k+1}$:

$$\mathbf{c}_{k+1} = \{\langle\langle x_{k+1}, c_k\rangle, \psi(c_{k+1})\rangle\}, \text{ where } \psi(c_{k+1}) = f(\mathbf{x}, \mathbf{c}_k).$$

[8] The existence of successive applications of merge (i.e., successive composition of f) leading to stable structures is assumed but it is worth to remark that a very restrictive set of relations K could lead us to the non existence of stable structure. The study of the combinatorial conditions needed to ensure the existence of an unbounded number of stable structures is not the objective of this work. In what follows, we are going to assume that compatibility relations are rich enough to enable the emergence of an unbounded number of stable structures.

The *state within the structure* of $\mathbf{c}_k$ and $\mathbf{x}_{k+1}$ that must be taken into account if either $\mathbf{x}_{k+1}$ or $\mathbf{c}_k$ are internally merged in later computations is

$$f(\mathbf{x}, \mathbf{c}_k) \text{ and } f(\mathbf{c}_k, \mathbf{x}),$$

respectively.

With the apparatus above defined, it is straightforward to generalize the role of feature checking to nests built using more than a single nesting machine. Since nothing qualitatively new is introduced, we left this part for the sake of clarity to enable the reader to focus into the more interesting forthcoming sections.

B. Three kinds of grammar

Let us study the kind of languages emerging when we have an alphabet $\mathbf{A}$, merge operation and a defined structure of features and compatibility relations Φ constraining the application of merge. The first key point relates to the decidibility problem. Whereas in the above sections we only looked at the structure to decide wether a given element was a sentence of the language we are generating, now we also look at the states of the elements. In general, the machinery equipped with an structure of features and compatibility relations Φ generate languages whose elements are stable syntactic objects. The generic symbol to refer to these languages is $\mathcal{L}$. This crucial element rules out many nests which were members of $\mathcal{F}_\mathcal{L}$ but not $\mathcal{L}$. Therefore, we can generally assume, except in trivial cases, that

$$\mathcal{L} \subset \mathcal{F}_\mathcal{L} \subset \mathcal{F}^*_\mathcal{L}.$$

As we shall see, we cannot assume that $|\mathcal{L}| = \infty$ in the general case. It is worth to note that the structures of a given language $\mathcal{L}$ are ready to be mapped into a semantic interpreter, like standard formal semantics. The role played by feature checking within the generation of structures of $\mathcal{L}$ is summarized in three qualitatively different behaviors. We define 3 types of *grammars*, according to the degree of restriction of the properties of the structure of features and compatibility relations Φ imposed when applying the merge operation:

1. A grammar is called *restrictive* if a given constituent γ is obtained through the merge operation of α, β in such a way that:

$$\gamma = \{\langle\langle\alpha,\beta\rangle, f(\alpha,\beta)\rangle\} \leftrightarrow (\exists i \leq k)(\langle\varphi_i(\alpha), \varphi_i(\beta)\rangle \in K_i)$$

   i.e., merge is only possible if there is at least one compatibility relation among the features of α and β.

2. A grammar is called *neutral/restrictive* if a given constituent γ is obtained through the merge operation of α, β in such a way that

$$\begin{aligned}\gamma \;&=\; \{\langle\langle\alpha,\beta\rangle, f(\alpha,\beta)\rangle\} \leftrightarrow \\ &\leftrightarrow\; [(\exists i \leq k)(\langle\varphi_i(\alpha),\varphi_i(\beta)\rangle \in K_i) \vee \\ &\quad \vee(\forall\varphi_j(x),\varphi_j(y))(\varphi_j(x) = 0 \vee \varphi_j(y) = 0)]\end{aligned}$$

i.e., merge is only possible if there is at least either a compatibility relation or the two elements are mutually neutral.

3. A grammar is called *non-restrictive* if merge is possible whatever the feature relations among elements.

Notice that *non-restrictive* grammars do not rule out the role of features, i.e, these grammars are not equivalent to the ones described in section II, where only structural information was taken into account. Indeed, the restriction that the state of the members of the language generated by such a grammar must be neutral clearly differentiates a non-restrictive grammar from the grammars where only the structure is taken into account. We also add that such a classification is not intended to be *exhaustive* but tries to be *reasonable*. Indeed, one can build a grammar which, depending on the elements that are defining a syntactic object, sometimes is restrictive or neutral/restrictive and sometimes non-restrictive, for example. Although it is possible we think that it introduces artifacts and that the proposed classification is the simplest one.

V. NESTING GRAMMARS AND THE CHOMSKY HIERARCHY

All the above discussed ingredients allow us to define the mathematical backbone of theoretical syntax. We do it through an unifying concept, namely the *nesting grammar*. A nesting grammar is a grammar including the formalism we proposed above. This means that, at the structural level, it generates nested structures by the successive application of merge over the alphabet elements or over previously formed nests. Furthermore, memory limitations restrict the number of nesting machines simultaneously at work and the access to the generated structures -cyclicity is, therefore, expected to occur. Additionally, a structure of features and compatibility relations is assumed.

Once such grammars are properly defined, one might ask about its generative power, depending on the memory bounds, specific properties of Φ and how this latter issue constrains the structure generation process. The end of this section is devoted to the derivation of the minimal conditions required for a grammar to have the power of a i)regular language and ii)context free language. The much more interesting -and also much more complex- case of context sensitive language is stated as a conjecture, but no rigorous proof has been yet achieved. It is worth to note that this latter point could shed light to an old conjecture stated by Noam Chomsky in the late fifties (Chomsky, 1957), where it was claimed that the grammar describing human language is *at least* a context-sensitive grammar (Chomsky, 1956; Hopcroft and Ullman, 1979).

A. Nesting grammars

Collecting all the ingredients developed in the previous pages, we have a compact definition of *nesting grammar*. A *nesting grammar* is the set of rules emerging from a set of nesting machines working in parallel, having finite memory and a structure of features and compatibility relations. We refer to a nesting grammar as G, which is described by a tuple:

$$G = \langle \mathbf{A}, \{D_1, ..., D_\mu\}, \Phi, \rho, \delta_c \rangle, \tag{13}$$

where

1. $\mathbf{A}$ is the alphabet, which is a finite set of singletons containing the element of A and its features. These are the *bricks* of syntactic structures -see eq. (10).

2. $D_1, ..., D_\mu$ are the nesting machine- that can simultaneously -but with a finite buffer of memory- participate in the generation of nests -see eq. (4).

3. Φ is the structure of features and compatibility relations -see eq. (12).

4. ρ can be either $r, n/r$ or n, and refers to the way the grammar applies the feature checking, either restrictive (r), neutral/restrictive (n/r) or non-restrictive (n) -see section IV.B.

5. δ_c is the upper bound on the distance of constituents that can be internally merged to generate another constituent -see eq. (7).

G will generate a language $\mathcal{L}$ whose elements are the nests generated under the conditions imposed by the nesting grammar G and such that:

$$\psi(N) = \overbrace{\langle 0, ..., 0 \rangle}^{k}$$

i.e., neutral nests.

The definition of nesting grammar organizes in a rigorous way the fundamental traits of theoretical syntax. Indeed, the nesting grammar is the theoretical object that would underly the theory of syntax, being its formal backbone. We observe that the framework is completely general and accepts a very rich variety of different options. For example, it can be used to encode a kind of grammar close to Montague's formal semantics (Montague and Thomason, 1974), with the difference that no assumptions on the semantic value are needed. It is intended to be, also, minimal, for it can be seen that many properties can be rigorously derived from it. Finally, we emphasize that it has the healthy property of being completely self consistent and that generates structures that can be mapped to feed a formal theory of how meaning is carried and compositionally generated.

B. Minimal properties of G and its location within the Chomsky Hierarchy

One of the most interesting consequences of the classification of three types of nesting grammars is its impact on the potential length and complexity of the generated structures. Indeed, the different parameters of working memory (μ, δ_c) and the structure of features and compatibility relations Φ can determine the complexity of a given nesting grammar. In this section we provide the minimal conditions for μ, δ_c, Φ to define a grammar located at a given level of the Chomsky Hierarchy. This connection of a framework based on merge operation and classical results of computation theory is the bridge that makes the amount of results on abstract grammars (Hopcroft and Ullman, 1979) potentially available for the study of syntax. We present it in a formal way, through two lemmas and one conjecture.

Lemma 2 *A Grammar generating a finite language. A grammar G such that*

$$G = \langle \mathbf{A}, \{D_1\}, \{\langle \varphi_1, K_1 \rangle\}, r, \delta_c \rangle$$

i.e., a restrictive grammar where

$$|\Phi| = |\{\langle \varphi_1, K_1 \rangle\}| = 1 \; \delta_c = 0.$$

will generate a language such that:

$$(\exists n \in \mathbb{N}) : |\mathcal{L}| < n,$$

(*a finite language.*)

Proof. If the grammar is *restrictive*, then every element $\{\mathbf{x}\} \in \mathbf{A}$ -notice that $\mathbf{A}$ is always finite- will be able to merge to those $\{\mathbf{y}\} \in \mathbf{A}$ such that $f(y, x) = 0$. Let F_x be the set of all members $\{\mathbf{y}\} \in \mathbf{A}$ such that $f(y, x) = 0$, namely,

$$F_x = \{\{y\} \in A : f(y, x) = 0\}$$

Clearly, given the finite nature of $\mathbf{A}$, $|F_x|$ is finite. Therefore, the constituent they built by merge, will be:

$$\mathbf{c}_1 = \langle \{\{x\}, \{x, y\}\}, f(x, y) \rangle = \langle \{\{x\}, \{x, y\}\}, 0 \rangle.$$

Since the grammar is restrictive, $\mathbf{c}_1$ will not be able to be merged to any bigger structure. Therefore, if we choose n such that

$$n = \sum_{x \in A} |F_x| + 1$$

(which is a finite number), then

$$|\mathcal{L}| = \sum_{x \in A} |F_x| < n,$$

thereby demonstrating the lemma.□

Remark. This implies that a grammar having the above described properties cannot be postulated as the grammar for a language by which

$$|\mathcal{L}| = \infty.$$

This point, even trivial, is crucial to emphasize, because human language is supposed to be potentially infinite. It is straightforward to realize that, if the grammar is neutral/restrictive, then unbounded structures can emerge, in spite that $|Phi| = 1$. In the same way, it is also easy to imagine that, if $|\Phi| > 1$ then, there are grammars such that, even restrictive, can generate unbounded structures. This crucial result must impact into the consideration of the role of feature checking in the theory of grammars.

Lemma 3. *Generation of a Context-free grammar. There is a grammar G such that*

$$G = \langle \mathbf{A}, \{D_1\}, \{\langle \varphi_1, K_1 \rangle\}, n/r, \delta_c \rangle$$

i.e., a neutral restrictive grammar with

$$|\Phi| = 1, \; \delta_c = 0, \; \mu = 1,$$

which is a context-free grammar.

Proof. Let us define the set of possible states that a constituent of $\mathcal{L}$ subject to the grammar defined in the lemma. Among the set of possible constituents, it defines an set of equivalence classes, to be named $\Psi(G)$, namely:

$$\Psi(G) = \{\Psi_1, ..., \Psi_m\}.$$

(In which the neutral state is included). We observe that, given the properties attributed to f, $\Psi(G)$ is finite. Therefore, all constituents can be classified as members of some equivalence class $\Psi_1, ...\Psi_m$. Now we define a rewriting grammar as follows:

$$\Psi_i \rightarrow \Psi_k \Psi_j \tag{14}$$

if:

$$f(\Psi_k, \Psi_j) = \Psi_i$$

and states Ψ_k and Ψ_j are neutral or compatible. We observe that using rewriting rules of the type (14) we can reproduce the behavior of G, thereby generating $\mathcal{L}$ (Hopcroft and Ullman, 1979). Furthermore, we observe that, since the left-hand of the set of the rewriting rules contains a single, non terminal symbol, and that $\Psi_1, ..., \Psi_m$ can refer to the states of both constituents and members of the alphabet $\mathbf{A}$, then they define a *context free grammar*, as we wanted to demonstrate. □

The next step would be to give the minimal relations to enable the emergence of a context-sensitive grammar, a much more complex and interesting object. However, up to now we state it as a conjecture.

Conjecture. *Generation of a Context-sensitive grammar. There is a grammar G such that*

$$G = \langle \mathbf{A}, \{D_1, D_2\}, \{\langle \varphi_1, K_1 \rangle\}, n/r, \delta_c \rangle$$

i.e., a neutral restrictive grammar with

$$|\Phi| = 1, \;\; \mu = 2, \; \delta_c > 0$$

which is generate a context-sensitive grammar.

The non-vanishing memory terms suggest that cyclicity would be possible and that it would be the source of non-adjacent relations enabling the jump to this stage of the Chomsky hierarchy. The demonstration of this conjecture would enable human language to better located within the framework of computation theory.

VI. DISCUSSION

In this paper we have shown that a rigorous backbone of theoretical syntax can be defined using the framework provided by set theory. Specially relevant for our study is *order theory*, which underlies almost all mathematical developments we presented above. The presented work is radically minimalist, for it reduces core syntax to set relations, deriving several properties from a very few principles. These organization in a few principles and its location within the hallmark of set theory has the interesting feature that the theory is well organized and no redundancies, vagueness or inaccuracies are expected to arise. Specifically, it defines a rigorous proposal to deal with the term *recursion*, which has been object of intense debates among scholars. It is our opinion that recursion has been often misunderstood, mainly to the fact that *recursive sets* have been at the core of modern mathematics, and it is a concept that goes beyond the intuitive idea of recursion in human language. Therefore, we claim that the concept of *nesting* clarifies this debate, providing a rigorous concept to deal with the intuitive idea of "recursion". More clarifications are needed, for example, concerning the "infinite recursion" attributed to human language: All structures of the nesting grammars, actual or potential, are finite. The size is unbounded, but finite, and, consistently, it is the number of potential sentences that is infinite.

The organization of the theory defines 3 different levels: The first one defines the way by which structures are generated, which results in the axiomatic definition of nest as the core structural concept. We named the set of potential structures as $\mathcal{F}^*_{\mathcal{L}}$. However, the intrinsic finite nature of the machinery generating nests introduces a second level, constraining the set of structures. Indeed, finite memory must be introduced in force, for it is clear that not to assuming it will lead us to the paradoxical situation that in a natural object there is an infinite mechanism of information storage. This second level of the theory describes the kind of nests that can be conceived when some bound on the memory of the machine is assumed, a set of structures that we named $\mathcal{F}_{\mathcal{L}}$. This finite memory condition has, as a result, the theoretical prediction that cyclic movement is expected to occur, and need not be postulated. We finally introduce a third constraint, which is not structural, but imposes restrictions on the merging operation depending on the labels of the sets to be merged. All three levels of analysis lead us to the definition of *nesting grammar*, the kind of theoretical object that we propose to underlie theoretical syntax. The construct can be considered, in some aspects, close to Montague's[9] program for grammar (Montague and Thomason, 1974) and it is worth to note that a nesting grammar could embed an important part of formal semantics. There are, however, important differences. The first is that, whereas Montague's grammar postulates *some generative mechanism* which is not considered the central issue of the theory, in the presented framework the way by which structures are formed occupies a central role, and thereby some consequences due to this mechanism cannot be observed in formal semantics. Beyond the central role played by the generation of naked structures, the syntax above proposed is much more flexible and it is not tied to any specific semantic theory, and richer versions than the one proposed by Montague could be considered.

The presented development can be reformulated in terms of axioms, thereby defining a legitimate schema for syntax to be included in the set of formal sciences, like physics. A tentative list, without trying to be definitive, would include 7 axioms underlying theoretical syntax. These axioms would be:

1. There is a finite alphabet of unstructured singletons $A = \{\{a_1\}, ..., \{a_n\}\}$
2. For every element of A we define a tuple of *features*, thereby generating the set $\mathbf{A}$, by which $(\forall \mathbf{x} \in \mathbf{A})$

   $$\mathbf{x} = \langle x, \langle \varphi_1(x), ..., \varphi_k(x) \rangle \rangle, \quad x \in A,$$

   where $\varphi_1, ..., \varphi(x)$ are functions $\varphi_i : A \to \mathbb{N}$, which encode the *features of* $\mathbf{x}$.
3. Merge is the fundamental operation and is *identified* as set union of alphabet elements or previously formed sets from alphabet elements.
4. Syntactic constituents are nests of arbitrary size obtained through successive applications of merge[10].
5. Merge is constrained by a set of compatibility relations K among features.
6. The *state* of a syntactic object is obtained by the application of a *checking function* along the derivation.

---

[9] The two e, t labels of Montague's grammar and the way by which they encode the compatibility in a given derivation could be perfectly encoded by natural numbers and a set of compatibility relations. Furthermore, the neutral assumption for well-formed elements of a given language would be analogous to the condition in which propositions are constructions where the truth value can be compositionally evaluated.

[10] We observe that this rules out the need to postulate the *single root condition*, since, in a nest

$$\exists \alpha : \alpha = \mathbf{larg}(\mathbf{N}).$$

7. The elements of the language are *neutral* with respect the compatibility relations:

$$(a \in \mathcal{L}) \Leftrightarrow (\psi(a) = \langle 0, .., 0 \rangle).$$

We observe that there is no mention to memory: Finiteness is not something that has to be imposed, but it is intrinsically attributed to the physical embodiment of the machinery generating the studied object. The converse option, not to assume finiteness, would need more justification. Further work would address the specific definition of a list of consistent axioms for the theoretical syntax.

How to deal with human language? We described a mechanism that generates a given language $\mathcal{L}$. Human syntax would accommodate to the framework of nesting grammar described above. This is the backbone, not the complete theory. The theory identifies how the formal issues are organized and which terms have to be considered under scrutiny. The elements of natural language would be *neutral nests formed by merge*. Such structures are *ready to be compositionally interpreted* by some semantic interpreter. In formal terms, the backbone of theoretical syntax for human language would be defined by the following nesting grammar G_{HL}:

$$G_{HL} = \langle \mathbf{Lex}, \{D_1, ..., D_\mu\}, \Phi, n/r, \delta_c \rangle. \qquad (15)$$

In this grammar, **Lex** would be the lexicon of a given language[11]. $D_1, ..., D_\mu$ the number of possible parallel nesting machines that can operate at the same time, which is a highly abstract object whose *observation* in real language can be obscured by many factors, since the observation of natural language cannot isolate the syntactic module. Φ is the structure of features and compatibility relations, which one can suppose to be about

$$|\Phi| \approx 2,$$

being the two kind of features the so-called *formal features* (agreement, etc...) and the so-called *semantic features*, (thematic relations, etc). The kind of application for the compatibility relations would be neutral/restrictive, for it is flexible enough but restricts some applications of merge. Furthermore, standard theories agree that *when there are no features to check* this structure belongs to the language and can be semantically interpreted. Finally, we consider that there are strong reasons to assume that:

$$\delta_c > 1,$$

therefore, internal merge or cyclic movement is expected to occur.

A final observation concerning human language refers the transition from two-words stage, to adult syntax. In the two words stage, syntactic structures are formed by two words having complementary semantic features, like $\langle \text{verb}, \text{noun} \rangle$. These semantic constraints are applied in a restrictive way, for it is uncommon to find structures like $\langle \text{verb}, \text{adjective} \rangle$, $\langle \text{noun}, \text{noun} \rangle$, etc. Therefore, we can consider that at the two words stage, we have a grammar of the type:

$$\langle \mathbf{Lex}, \{D_1\}, \{\langle \varphi_1, K_1 \rangle\}, r, \delta_c \rangle$$

(where δ_c could be either 0 or 1). This grammar generates a finite language, as shown in lemma 3. Thus, the jump to an adult grammar as the one proposed above - see the definition of G_{HL} in equation (15) would imply a transition from a finite-state grammar to at least, context-free grammar -but we conjectured that the jump might be sharper, to a context -sensitive language.

The brief discussion provided above concerning the transition to adult syntax is a nice example of how this formalism organizes and properly describes syntactic phenomena. However, we provided only the first *draft* of how the theoretical syntax might be organized. There is still hard work to properly identify how features actually work, or what is the role of the interfaces. We encourage researchers to do it with the aim of being consistent both with the adjacent theories -in this case, formal semantics, in one side, and natural computation, on the other side- and internally. We proposed ordered set theory as the formal apparatus where to locate theoretical syntax. The underlying philosophy might not be different to the choice made by quantum mechanics by choosing Hilbert spaces in the 30's or general relativity, which chose differential geometry to provide a solid background for the theory.

[11] I do not go into the discussion on how differentiated is the morphology from the syntax.

Acknowledgments

The author acknowledges the helpful comments made by Jordi Fortuny in earlier versions of the manuscript. This work has been founded by the James McDonnell Foundation.

References

Chomsky, N., 1956, IEEE Transactions on Information Theory **2**, 113.

Chomsky, N., 1957, *Syntactic Structures* (The Hague. Mouton).

Chomsky, N., 1995, *The minimalist program* (MIT Press).

Chomsky, N., 2008, Foundational issues in linguistic theory: Essays in honor of Jean-Roger Vergnaud , 133.

Fortuny, J., and B. Corominas-Murtra, 2009, Catalan Journal of Linguistics **8**, 99.

Hauser, M. D., N. Chomsky, and T. W. Fitch, 2002, Science **298**, 1569.

Hopcroft, J., and J. Ullman, 1979, *Introduction to Automata Theory, Languages and Computation* (Addison-Wesley. New York).

Hornstein, N., 2000, *Move! A minimalist Theory of Construal* (Blackwell, Oxford).

Kelley, J., 1955, *General Topology* (Van Nostrand).

Kuratowski, K., 1921, Fundamenta Methematicae **2**, 161.

Montague, R., and R. Thomason, 1974, *Formal philosophy: selected papers of Richard Montague* (Yale University Press).

Partee, B., A. Ter Meulen, and R. Wall, 1990, *Mathematical methods in linguistics* (Springer).

Radford, A., 1997, *Syntax: A minimalist introduction* (Cambridge University Press. Cambridge).

Rosselló, J., 2006, in *The Biolinguistic Turn. Issues on Language and Biology*, edited by J. Rosselló and T. Martín (Promociones y Publicaciones Universitarias, S.A.), pp. 162–186.

Rosselló, J., 2008, *Introducció a la sintaxi* (Ms, Universitat de Barcelona).

Smith, J. M., and E. Szathmáry, 1995, *The Major Transitions In Evolution* (Oxford University Press).

Suppes, P., 1972, *Axiomatic Set Theory* (Dover: New York).

Uriagereka, J., 2002, *Derivations: Exploring the Dynamics of Syntax* (Routledge).

Chapter 7

PAPERS ON ZIPF'S LAW

Corominas-Murtra B, Solé RV. Universality of Zipf's law. Phys Rev E Stat Nonlin Soft Matter Phys. 2010; 82(1 Pt1): 011102.

Corominas-Murtra B, Fortuny J, Solé RV. Emergence of Zipf's law in the evolution of communication. Phys Rev E Stat Nonlin Soft Matter Phys. 2011; 83(3 Pt 2): 036115.

Chapter 8

PAPERS ON REFERENTIALITY

Corominas B, Solé RV. Network topology and self-consistency in language games. J Theor Biol. 2006; 241(2): 438-41.

Corominas-Murtra B, Fortuny Andreu J, Solé R. Coding and decording in the evolution of communication: information richness and referentiality. arxiV.org. Cornell University Library. Consultat: 13 set. 2011. Disponible a : arXiv:1004.1999v1

Coding and decoding in the Evolution of communication: Information richness and Referentiality

Bernat Corominas Murtra,[1] Jordi Fortuny Andreu,[2] and Ricard Solé[1, 3, 4]
[1]ICREA-Complex Systems Lab, Universitat Pompeu Fabra (GRIB), Dr Aiguader 80, 08003 Barcelona, Spain
[2]Centre de Lingüística Teòrica (CLT), Facultat de Lletres, Edifici B, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain
[3]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM 87501, USA
[4]Institut de Biologia Evolutiva. CSIC-UPF. Passeig Maritim de la Barceloneta, 37-49, 08003 Barcelona, Spain.

One of the most basic properties of the communicative sign is its dual nature. That is, a sign is a twofold entity composed of a formal component, which we call *signal*, and a referential component, namely a *reference*. Based on this conception, we say that a referent is coded in a particular sign, or that a sign is decoded in a particular referent. In selective scenarios it is crucial for the success of any adaptive innovation or communicative exchange that, if a particular referent a is coded in a particular signal s during the coding process, then the referent a is decoded from the sign s during the decoding process. In other words the *referentiality of a signal* must be preserved after being decoded, due to a selective pressure. Despite the information-theoretic flavour of this requirement, an inquiry into classical concepts of information theory such as entropy or mutual information will lead us to the conclusion that information theory as usually stated does not account for this very important requirement that natural communication systems must satisfy. Motivated by the relevance of the preservation of referentiality in evolution, we will fill this gap from a theoretical viewpoint, by deriving the consistent information conveyed from an arbitrary coding agent A^v to an arbitrary decoding agent A^u and discussing several of its interesting properties.

Keywords: entropy, information, referentiality, consistent information

I. INTRODUCTION

Biological Systems store and process information at many different scales (Yockey, 1992). Organisms or cells react to changes in the external environment by gathering information and making the right decisions -once such information is properly interpreted. In a way, we can identify the external changes as input signals to be coded and decoded by the cellullar machinery or information processing of neural networks, and include the exchange of signals between individuals or abstract agents sharing a given communication system (Hurford, 1989; Komarova and Niyogi, 2004; Niyogi, 2006; Nowak and Krakauer, 1999).

The ability to store information to interpret the surroundings beyond pure noise is thus an important property of biological systems. An organism or abstract agent can make use of this feature to react to the environment in a selectively advantageous way. This is possible provided that, in biological systems, a communicative signal must be necessarily linked to a referential value, that is, it must have a meaningful content. As pointed out by John Hopfield:

> *Meaningful content, as distinct from noise entropy, can be distinguished by the fact that a change in a meaningful bit will have an effect on the macroscopic behavior of a system* (Hopfield, 1994).

The meaningful content of information can be understood as something additional to classical information which is preserved through generations (or by the members of a given population in a given communicative exchange) resulting in a *consistent* response to the environment (Haken, 1978).

The explicit incorporation of the referential value in the information content is, in some sense, external to classical information theory, since, roughly speaking, the standard measure of mutual information only accounts for the relevance of correlations among sets of random variables. Indeed, one can establish configurations among coder and decoder by which mutual information is maximal but the referentiality value of the signal is lost during the communicative exchange. Let us consider the following example: Suppose a system where the event *fire* is coded as the signal a, and that such a signal a is always decoded as the event *water*. Suppose, also, that the event *water* is coded as the signal b and it is always decoded as *fire*. In this system, both the coder and the decoder depict a one-to-one mapping between input and output, and the mutual information between the set of events shared by coder and decoder would be maximum. However, if we take the system as a whole, the non-preservation of any referential value renders the communication code useless.

Not surprisingly, evolutionary experiments involving artificial agents (such as robots) include, as part of the selective pressures, the consistency of signals and referents. If survival or higher scores depend on a fitness measure which requires a proper sharing of information, the final outcome of the dynamics is a set of agents using common signals to refer to the same object (Nolfi and Mirolli, 2010; Steels, 2001; Steels and Baillie, 2003). Formally, we say that the communicative sign has a dual

nature[1]: a sign would involve a pair

$$\langle m_i, s_k \rangle, \tag{1}$$

composed of a *signal*, s_i, and a *referent*, m_k. Such pair must be conserved in a consistent communicative interchange.

The problem of consistency of the communicative process was early addressed in (Hurford, 1989), through a formalism consisting in signal/referent matrices. Further works showed the suitability of such formalism, and enabled the study of the emergence of consensus driven by selective forces (Nowak and Krakauer, 1999). These studies showed that an evolutionary process could result in a shared code by a population of interacting agents. Under this framework, the existence of optimal solutions has been studied (Komarova and Niyogi, 2004), as well as the problem of the information catastrophe or *linguistic error limit* (Nowak, 2000), using evolutionary game theory involving a payoff function accounting for the average number of well-referentiated signals.

It is the purpose of this theoretical work to rigorously identify the amount of information which conserves the dual structure of a sign, i.e., the amount of *consistent information*, and to explore some of its consequences. Specifically, we evaluate the relevance of the consistent input/output pairs, assuming that the input set and the output set are equal. The study of the behaviour of the consistent information displays interesting differences with classical Shannon's mutual information.

We should properly differentiate the problem of consistency from the problem of *absolute information content* of a given signal -or, in general, mathematical object. The latter arises from the fact that, in Shannon's information theory, the information content of a given signal is computed from the relative abundance of such a signal against the occurrences of the whole set of signals. The information content of an isolated signal is not defined (or equal to zero). This is solved by the definition of the Kolmogorov Complexity (Cover and Thomas, 1991; Kolmogorov, 1965; Ming and Vitányi, 1997), which can be understood as the absolute information content of a given signal -or mathematical object. Our purpose can be embedded in Shannon's framework. Accepting the relative nature of the information content, we attack the problem of the consistency of input/output pairs.

The paper is written in a self-contained way. Thus, beyond basics of probability theory we properly introduce the concepts and the required mathematical apparatus. At the end of the paper, a case study (the classical binary symmetric channel) is described in detail.

[1] This central property of the communicative sign resembles the *duality of the linguistic sign* pointed out by first time by the Swiss linguist Ferdinand de Saussure (Saussure, 1916). According to Saussure, a linguistic sign is a psychical unit with two faces: a signifier and a signified. The former term is close to our term 'signal' and the latter to our term 'reference'. There are, though, important differences between the information-theoretical approach we are about to develop and Saussure's conception of the linguistic sign.

II. THE MINIMAL SYSTEM AND ITS ASSOCIATED INFORMATION MEASURES

In this section we define the minimal system composed of two agents able to both code and decode a set of external events.

A. The communicative system

Consider a set of (at least, two) interacting agents *living* in a shared world (Komarova and Niyogi, 2004). Agents communicatively interact through noisy channels. The description of this system is based on the probability transition matrices defining the coding and decoding processes, the probability transition matrix for the channel and the random variables associated to the inputs and outputs, which account for the successive information processing through the system formed by two agents and the noisy channel -see fig.1. The qualitative difference with respect to the classical communication scheme is that we take into account the particular value of the input and the output therby capturing the referential value of the communicative exchange. An *agent*, A^v, is defined as a pair of computing devices,

$$A^v \equiv \{\mathbf{P}^v, \mathbf{Q}^v\}, \tag{2}$$

where $\mathbf{P}^v$ is the coder module and $\mathbf{Q}^v$ is the decoder module. The shared world is defined by a random variable X_Ω which takes values on the set of events, Ω:

$$\Omega = \{m_1, ..., m_n\}, \tag{3}$$

being the (always non-zero) probability associated to any event $m_k \in \Omega$ defined by $\mu(m_k)$. The coder module, $\mathbf{P}^v$, is described by a mapping from Ω to the set $\mathcal{S}$:

$$\mathcal{S} = \{s_1, ..., s_n\}, \tag{4}$$

to be identified as the set of signals. For simplicity, here we assume $|\Omega| = |\mathcal{S}| = n$. This mapping is realized according to the following matrix of transition probabilities:

$$\mathbf{P}^v_{ij} = \mathbb{P}_v(s_j|m_i), \tag{5}$$

which satisfies the following condition:

$$(\forall m_i \in \Omega) \ \sum_{j \leq n} \mathbf{P}^v_{ij} = 1. \tag{6}$$

The output of the coding process is described by the random variable X_s, taking values on $\mathcal{S}$ according to the probability distribution ν:

$$\nu(s_i) = \sum_{k \leq n} \mu(m_k)\mathbf{P}^v_{ki}. \tag{7}$$

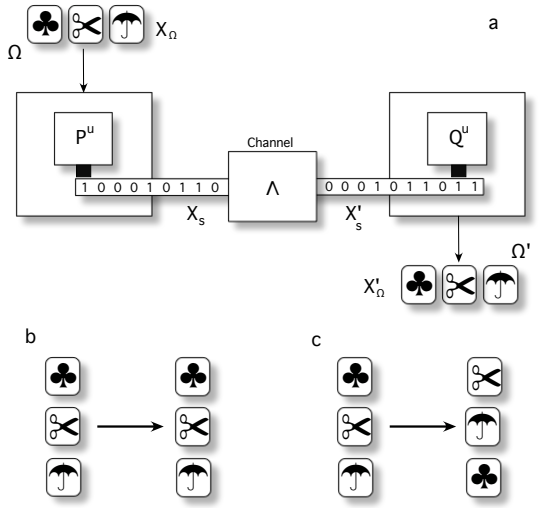


FIG. 1 The minimal communicative system to study the conservation of referentiality (a): A shared world, whose events are the members of the set Ω and whose behavior is governed by the random variable X_Ω. A coding engine, $\mathbf{P}^v$, which performs a mapping between Ω and the set of signals $\mathcal{S}$, being X_s the random variable describing the behavior of the set of signals obtained after coding. The channel, Λ, may be noisy and, thus, the input of the decoding device, $\mathbf{Q}^v$, depicted by X'_s, might be different from X_s. $\mathbf{Q}^v$ performs a mapping among $\mathcal{S}$ and Ω whose output is described by X'_Ω. Whereas the mutual information provides us a measure of the relevance of the correlations among X_Ω and X'_Ω, the *consistent information* evaluates the relevance of the information provided by consistent pairs on the overall amount of information. In this context, from a pure information-theoretical point of view, situations like b) and c) could be indistinguishable. By defining the so-called consistent information we can properly differentiate b) and c) by evaluating the degree of consistency of input/output pairs -see text.

The channel, Λ, is characterized by the $n \times n$ matrix of conditional probabilities $\mathbb{P}_\Lambda(\mathcal{S}|\mathcal{S})$, i.e.,

$$\Lambda_{ij} = \mathbb{P}_\Lambda(s_j|s_i). \tag{8}$$

The output of the composite system coder+channel, $\mathbf{P}^v\Lambda$, is described by the random variable X'_s, which takes values on the set $\mathcal{S}$ following the probability distribution ν', defined as:

$$\nu'(s_i) = \sum_k \mu(m_k)\mathbb{P}_{v\Lambda}(s_i|m_k), \tag{9}$$

where

$$\mathbb{P}_{v\Lambda}(s_i|m_k) = \sum_{j\leq n} \mathbf{P}^v_{kj}\Lambda_{ji}. \tag{10}$$

Finally, the decoder module is a computational device described by a mapping from $\mathcal{S}$ to Ω, i.e it receives $\mathcal{S}$ as the input set, emitted by another agent through the channel, and yields as output elements of the set Ω. $\mathbf{Q}^v$ is completely defined by its transition probabilities, i.e.:

$$\mathbf{Q}^v_{ik} = \mathbb{P}_v(m_k|s_i), \tag{11}$$

which satisfies the following condition:

$$(\forall s_i \in \mathcal{S}) \ \sum_{k\leq n} \mathbf{Q}^v_{ik} = 1. \tag{12}$$

Aditionally, we can impose another condition:

$$(\forall m_j \in \Omega) \ \sum_{i\leq n} \mathbb{P}_v(s_i|m_j) = 1, \tag{13}$$

which is necessary for A^v to reconstruct Ω, i.e., if the population of interacting agents share the world. By imposing condition (13) we avoid configurations in which some $m_k \in \Omega$ cannot be referentiated by the decoder agent. We notice that it is consistent with the fact that no element from Ω has zero probability to occur. Furthermore, we emphasize the assumption that, in a given agent A^v, following (Nowak and Krakauer, 1999; Plotkin and Nowak, 2000) but not (Hurford, 1989; Komarova and Niyogi, 2004) there is a priori no correlation between $\mathbf{P}^v$ and $\mathbf{Q}^v$. Finally, under the presence of another agent A^u, we can define the output of $\mathbf{Q}^v$ as the random variable X'_Ω, taking values on the set Ω and following the probability distribution μ', which takes the form:

$$\mu'(m_i) \equiv \sum_{l\leq n} \mu(m_l)\mathbb{P}_{A^u\to A^v}(m_i|m_l), \tag{14}$$

where

$$\mathbb{P}_{A^u\to A^v}(m_i|m_l) \ = \ \sum_{j,r\leq n} \mathbf{P}^u_{lj}\Lambda_{jr}\mathbf{Q}^v_{ri}, \tag{15}$$

$$\mathbb{P}_{A^v\to A^u}(m_i,m_j) \ = \ \sum_{l,r} \mu(m_j)\mathbf{P}^v_{jl}\Lambda_{lr}\mathbf{Q}^u_{ri}. \tag{16}$$

Consistently,

$$\sum_{i,l\leq n} \mathbb{P}_{A^u\to A^v}(m_i|m_l) = n. \tag{17}$$

Once we have the description of the different pieces of the problem, we proceed to study the couplings among them in order to obtain a suitable measure of the consistency of the communicative process. The first natural quantitative observable to account for the degree of consistency is the fraction of events $m_i \in \Omega$ which are consistently decoded. From eq. (16) it is straigtforward to conclude that such a fraction ($F(A^v \to A^u)$) is given by:

$$F(A^v \to A^u) = \sum_{i\leq n} \mathbb{P}_{A^v\to A^u}(m_i,m_i). \tag{18}$$

And if we take into account that the communicative exchange takes place in both directions, we have:

$$F(A^v, A^u) = \frac{1}{2}\left(F(A^v \to A^u) + F(A^u \to A^v)\right). \quad (19)$$

Putting aside slight variations, eq. (19) has been widely used as a payoff function to study the emergence of consistent codes -in terms of duality preservation- through an evolutionary process involving several agents in every generation (Hurford, 1989; Komarova and Niyogi, 2004; Nowak and Krakauer, 1999; Plotkin and Nowak, 2000). Such an evolutionary dynamics yielded important results which help understanding how selective pressures push a population of communicating agents to reach a consensus in their internal codes.

B. Mutual Information

Now we proceed to compute the mutual information among relevant variables of the system. We stress that it does not account for the referentiality of the sent signals. Instead, it quantifies, in bits, the relevance of the correlations among two random variables, as a potential message conveyer system, never specifying the referential value of any sequence or signal.

Let us briefly review some fundamental definitions and concepts of information theory. We know that, given two random variables X, Y, with associated probability functions $p(x), p(y)$, conditional probabilities $\mathbb{P}(x|y), \mathbb{P}(y|x)$ and joint probabilities $\mathbb{P}(x, y)$, its mutual information $I(X : Y)$ is defined as (Ash, 1990; Cover and Thomas, 1991; Shannon, 1948):

$$I(X:Y) \;=\; \sum_{x,y} \mathbb{P}(x,y) \log \frac{\mathbb{P}(x,y)}{p(x)p(y)}, \quad (20)$$

or equivalently:

$$I(X:Y) = H(X) - H(X|Y), \quad (21)$$

being $H(X)$ the *Shannon entropy* or *uncertainty* associated to the random variable X:

$$H(X) = -\sum_x p(x) \log p(x), \quad (22)$$

and $H(X|Y)$ the *conditional entropy* or *conditional uncertainty* associated to the random variable X with respect to the random variable Y:

$$H(X|Y) = -\sum_y p(y) \sum_x \mathbb{P}(x|y) \log \mathbb{P}(x|y). \quad (23)$$

We can also define the *joint entropy* among two random variables X, Y, written as $H(X, Y)$:

$$H(X,Y) = -\sum_{x,y} \mathbb{P}(x,y) \log \mathbb{P}(x,y). \quad (24)$$

A key concept of information theory is the so-called *channel capacity*, $C(\Lambda)$, which, roughly speaking, is the maximum amount of bits that can be reliably processed by the system, namely:

$$C(\Lambda) = \max_{p(x)} I(X:Y). \quad (25)$$

As usual, in our minimal system of two interacting agents we explicitely introduced the channel, Λ, as a matrix of transition probabilities between the two agents. Channel capacity is an intrinsic feature of the channel; as the fundamental theorem of information theory (Ash, 1990; Cover and Thomas, 1991; Shannon, 1948) states, it is possible to send any message of R bits through the channel with an arbitrary small probability of error if:

$$R < C(\Lambda); \quad (26)$$

otherwise, the probability of errors in transmission is no longer negligible. One should not confuse the statements concerning the capacity of the channel with the fact that given a random variable with associated probability distribution $p(x)$, we have:

$$\max I(X:Y) = H(X) = H(Y) \quad (27)$$

(provided that $C(\Lambda) > H(X)$). In those cases, we refer to the channel as *noiseless*.

Let us now return to our system. Using eq. (20) and the joint probabilities derived in eq. (16), we can compute the mutual information among X_Ω and X'_Ω when A^v is the coder and A^u the decoder, to be noted $I(A^v \to A^u)$, as follows:

$$I(A^v \to A^u) \;=\; \sum_{j,i \leq n} \mathbb{P}_{A^v \to A^u}(m_i, m_i) \times \\ \times \log\left(\frac{\mathbb{P}_{A^v \to A^u}(m_i, m_j)}{\mu(m_i)\mu'(m_j)}\right). \quad (28)$$

Notice that, since the coding and decoding modules of a given agent are depicted by different, a priori non-related matrices, in general

$$I(A^v \to A^u) \neq I(A^u \to A^v). \quad (29)$$

The average of shared information among agent A^v and A^u will be:

$$\langle I(A^v, A^u)\rangle = \frac{1}{2}(I(A^v \to A^u) + I(A^u \to A^v)). \quad (30)$$

Clearly, since the channel is the same in both directions of the communicative exchange, the following inequality holds:

$$\langle I(A^v, A^u)\rangle < C(\Lambda). \quad (31)$$

In the next section we investigate the role of the *well-correlated pairs* and its impact in the overall quantity of information.

III. CONSISTENT INFORMATION

To obtain the amount of consistent information shared among A^u and A^v, we must find a special type of correlations among X_Ω and X'_Ω. Specifically, we are concerned with the observations of both coder and decoder such that the input and the output are the same element, i.e., the fraction of information that can be extracted from the observation of all consistent pairs $\mathbb{P}_{A^v\to A^u}(m_i, m_i)$. This fraction is captured by the so-called *referential parameter*, and its derivation is the objective of the next subsection.

A. The Referential parameter

The mutual information among two random variables is obtained by exploring the behavior of input/output pairs, averaging the logarithm of the relation among the actual probability to find a given pair and the one expected by chance. Consistently, the referential parameter is thus obtained by averaging the fraction of information that can be extracted by observing consistent pairs against the whole information we can obtain by looking at all possible ones.

1. Derivation of the Referential Parameter σ

Following the standard definitions of the information conveyed by a signal (Shannon, 1948), the information we extract from the observation of a pair input-output m_i, m_j is:

$$-\log \mathbb{P}_{A^v\to A^u}(m_i, m_j). \tag{32}$$

Following eq. (24), the average of information obtained from the observation of pairs will be precisely the joint entropy between X_Ω and X'_Ω, $H(X_\Omega, X'_\Omega)$:

$$-\sum_{i,j\leq n} \mathbb{P}_{A^v\to A^u}(m_i, m_j)\log \mathbb{P}_{A^v\to A^u}(m_i, m_j).$$

Let us simplify the notation by defining a matrix $\mathbf{J}$. The elements of such a matrix are the joint probabilities, namely:

$$J_{ij} \equiv \mathbb{P}_{A^v\to A^u}(m_i, m_j). \tag{33}$$

From the above matrix, we can identify the contributions of the consistent pairs by looking at the elements of the diagonal. The relative impact of consistent pairs on the overall measure of information will define the *referential parameter* associated to the communicative exchange $A^v \to A^u$, to be indicated as $\sigma_{A^v\to A^u}$. This is our key definition, and its explicit form will be:

$$\sigma_{A^v\to A^u} \equiv -\frac{\mathbf{tr}(\mathbf{J}\log\mathbf{J})}{H(X_\Omega, X'_\Omega)}, \tag{34}$$

where $\mathbf{tr}(\mathbf{J}\log\mathbf{J})$ is the *trace* of the matrix $\mathbf{J}\log\mathbf{J}$, i.e.:

$$\mathbf{tr}(\mathbf{J}\log\mathbf{J}) = \sum_{i\leq n} J_{ii}\log J_{ii}. \tag{35}$$

By dividing $\mathbf{tr}(\mathbf{J})$ by $H(X_\Omega, X'_\Omega)$ we capture the fraction of bits obtained from the observation of consistent pairs against all possible pairs $\langle m_i, m_j\rangle$[2].

The amount of *Consistent Information*, $\mathcal{I}(A^v \to A^u)$, is obtained by weighting the overall mutual information with the referential parameter:

$$\mathcal{I}(A^v \to A^u) = I(A^v \to A^u)\sigma_{A^v\to A^u}. \tag{36}$$

The average of consistent information among two agents, $\mathcal{F}(A^v, A^u)$ will be, consistently:

$$\mathcal{F}(A^v, A^u) \equiv \frac{1}{2}\left(\mathcal{I}(A^v \to A^u) + \mathcal{I}(A^v \to A^u)\right). \tag{37}$$

Since $\sigma_{A^v\to A^u} \in [0, 1]$, from the defintion of channel capacity and the symmetry properties of the mutual information, it is straightforward to show that:

$$\mathcal{F}(A^v, A^u) \leq \langle I(A^v, A^u)\rangle \leq \mathcal{C}(\Lambda).$$

Eqs. (34, 36) and (37) are the central equations of this paper. Let us focus on eq. (36). In this equation, we derive the average of consistent bits in a minimal system consisting of two agents (coder/decoder). Consistent information has been obtained by mathematically inserting the dual nature of the communicative sign - which forces the explicit presence of coder, channel and decoder modules- and subsequently selecting the subset of correlations by which the input symbol (the specific realization of X_Ω) is equal to the output symbol (i.e., the specific realization of X'_Ω). Eq. (37) accounts for the (possibly) symmetrical nature of the communicative exchange among agents: a priori, all agents can be both coder and decoder, and we have to evaluate and average the two possible configurations. The information-theoretic flavour of $\mathcal{F}$ enables us to study the conservation of referentiality from the well-grounded framework of Information Theory.

B. General Behavior of Consistent Information

So far we have been concerned with the derivation of the amount of information which is consistently decoded,

[2] We might notice that the amount of information carried by consistent pairs resembles the formal exposition of the Von Neumann entropy for quantum states, $S(\rho)$, which captures the degree of mixture of a given quantum state and its associated uncertainty in measuring (Von Neumann, 1936). In this way, we observe that S can be, roughly speaking, identified with an indicator of the consistency of the quantum state. However, it is worth noting that these measures are conceptually and formally different.

taking into account the dual nature of the communicative sign -equations (34), (36) and (37). Now we explore some of its properties, and we highlight the conceptual and quantitative differences between $\mathcal{I}$ and I.

To study the behavior of $\mathcal{I}$ and its relation to I, we will isolate the first three most salient features. Specifically, we shall concern ourselves with the following logical implications:

$$i)\ (\sigma_{A^v\to A^u}=1)\ \Rightarrow\ (I(A^v\to A^u)=H(X_\Omega)),\ (38)$$
$$ii)\ (\sigma_{A^v\to A^u}=1)\ \not\Leftarrow\ (I(A^v\to A^u)=H(X_\Omega)).\ (39)$$

The first i) implication refers to the perfect conservation of referentiality, which, in turn, implies maximum mutual information. However, the inverse, ii), is not generally true, since, as we shall see, there are many situations by which the mutual information is maximum although there is no conservation of referentiality. Furthermore, we consider a third case, the noisy channel (which implies that $H(X_\Omega|X'_\Omega)>0$). In this case:

$$iii)\ H(X_\Omega)>I(A^v\to A^u)>\mathcal{I}(A^v\to A^u).\quad (40)$$

We begin with the implications i) and ii). In both cases, the whole process is noiseless, since from eq. (27) $\max I(A^v\to A^u)=H(X_\Omega)$). To address the first logical implication, i), we obtain the typology of configurations of $\mathbf{P}^v,\Lambda,\mathbf{Q}^v$ leading to $\sigma_{A^v\to A^u}=1$. We observe that the condition (39) is achieved if $\mathbb{P}(X'_\Omega|X_\Omega)=\mathbb{1}$, i.e., the identity matrix:

$$\mathbb{1}_{ij}=\begin{cases}1 \text{ iff } i=j\\ 0 \text{ otherwise.}\end{cases}\quad (41)$$

Such a condition only holds if

$$\mathbf{P}^v=(\Lambda\mathbf{Q}^u)^{-1},\quad (42)$$

since given a square matrix $\mathbf{A}$, $\mathbf{A}\cdot\mathbf{A}^{-1}=\mathbb{1}$ -provided that $\mathbf{A}^{-1}$ exists. From the conditions imposed over the transition matrices provided in eqs. (6,12,17), the above relation is fullfilled if and only if all the matrices $\mathbf{P}^v,\Lambda,\mathbf{Q}^u$ are *permutation matrices*. Let us briefly revise this concept, which will be useful in the following lines. A *permutation matrix* is a square matrix which has exactly one entry equal to 1 in each row and each column and 0's elsewhere. For example, if $n=3$, we have 6 permutation matrices, namely:

$$\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix},\begin{pmatrix}1&0&0\\0&0&1\\0&1&0\end{pmatrix},\begin{pmatrix}0&1&0\\1&0&0\\0&0&1\end{pmatrix},\quad (43)$$

$$\begin{pmatrix}0&0&1\\1&0&0\\0&1&0\end{pmatrix},\begin{pmatrix}0&1&0\\0&0&1\\1&0&0\end{pmatrix},\begin{pmatrix}0&0&1\\0&1&0\\1&0&0\end{pmatrix}.\quad (44)$$

The set of $n\times n$ permutation matrices is idicated as $\Pi_{n\times n}$ and it can be shown that, if $\mathbf{A}\in\Pi_{n\times n}$, $\mathbf{A}^{-1}=\mathbf{A}^T\in\Pi_{n\times n}$ and, if $\mathbf{A},\mathbf{B}\in\Pi_{n\times n}$, the product $\mathbf{AB}\in\Pi_{n\times n}$. Furthermore, it is clear that $\mathbb{1}\in\Pi_{n\times n}$. If we translate the above facts of permutation matrices to our problem, we find that $\sigma_{A^v\to A^u}=1$ is achieved if:

$$(\mathbf{P}^v,\Lambda,\mathbf{Q}^u\in\Pi_{n\times n})\text{ and }\mathbf{P}^v=(\Lambda\mathbf{Q}^u)^T,\quad (45)$$

leading to the following chain of equalitites, which only holds in this special case:

$$\begin{aligned}\mathcal{I}(A^v\to A^u)&=I(A^v\to A^u)\\&=\max I(A^v\to A^u)\\&=H(X_\Omega).\end{aligned}$$

Case ii) is easily demonstrated by observing that, if $\mathbb{P}(X_\Omega|X'_\Omega)\in\Pi_{n\times n}$, then $\mathbb{P}(X'_\Omega|X_\Omega)\in\Pi_{n\times n}$ and thus

$$H(X_\Omega|X'_\Omega)=0,\quad (46)$$

leading to:

$$I(A^v\to A^u)=\max I(A^v\to A^u)=H(X_\Omega),\quad (47)$$

which is achieved only imposing that

$$\mathbf{P}^v,\Lambda,\mathbf{Q}^u\in\Pi_{n\times n}.\quad (48)$$

However, as we saw above, only a special configuration of permutation matrices leads to $\sigma_{A^v\to A^u}=1$. Thus, for the majority of cases where $I(A^v\to A^u)=\max I(A^v\to A^u)$, the conservation of the referentiality fails, leading to

$$I(A^v\to A^u)>\mathcal{I}(A^v\to A^u),\quad (49)$$

unless condition (45) is satisfied. Let us notice that there are limit cases where, although $I(A^v\to A^u)=\max I(A^v\to A^u)$, $\mathcal{I}(A^v\to A^u)=0$, since it is possible to find a configuration of $\mathbf{P}^v,\Lambda,\mathbf{Q}^u\in\Pi_{n\times n}$ such that $\mathbb{P}(X_\Omega|X'_\Omega)$ is a permutation matrix with all zeros in the main diagonal, leading to $\sigma_{A^v\to A^u}=0$.

Case iii) is by far the most interesting, since natural systems are noisy, and the conclusion could invalidate some results concerning the information measures related to systems where referentiality is important. The first inequality trivially derives from equation (21), from which we conclude that $I(A^v\to A^u)<H(X_\Omega)$. The argument to demonstrate the second inequality lies on the following implication:

$$(H(X_\Omega|X'_\Omega)>0)\Rightarrow(\mathbb{P}_{A^v\to A^u}(X'_\Omega|X_\Omega)\notin\Pi_{n\times n}).\quad (50)$$

Indeed, let us proceed by contradiction: Let us suppose that $\mathbb{P}_{A^v\to A^u}(X'_\Omega|X_\Omega)\in\Pi_{n\times n}$. Then, as discussed above, $\mathbb{P}_{A^v\to A^u}(X_\Omega|X'_\Omega)\in\Pi_{n\times n}$. But this should imply that $H(X_\Omega|X'_\Omega)=0$, thus contradicting the premise that $H(X_\Omega|X'_\Omega)>0$.

This has a direct consequence. Since such conditional probabilities satisfy eq. (17), then, more than n matrix elements of $\mathbb{P}_{A^v\to A^u}(X_\Omega|X'_\Omega)$ must be different from zero. The same applies to the matrix of joint probabilities $\mathbf{J}$

and thus it also applies to $-\mathbf{J}\log\mathbf{J}$. Since the trace is a sum of n elements, it should be clear that, under noise:

$$H(X_\Omega, X'_\Omega) > -\mathbf{tr}(\mathbf{J}\log\mathbf{J}), \tag{51}$$

leading to:

$$\sigma_{A^v\to A^u} < 1, \tag{52}$$

thus recovering the chain of inequalities provided in eq. (40):

$$H(X_\Omega) > I(A^v \to A^u) > \mathcal{I}(A^v \to A^u). \tag{53}$$

If we expand the reasoning to the symmetrical consistent information $\mathcal{F}(A^v, A^u)$ defined in (37):

$$\mathcal{F}(A^v, A^u) < \langle I(A^v, A^u)\rangle. \tag{54}$$

We see that referentiality conservation introduces an extra source of dissipation of information. In those scenarios where referentiality conservation is an important advantage, the dissipation of information, $\mathcal{I}_D$, among two agents has two components:

$$\mathcal{I}_D = \overbrace{H(X_\Omega|X'_\Omega)}^{physical\ noise} + \overbrace{(1-\sigma)I(A^v \to A^u)}^{Referential\ noise}, \tag{55}$$

being the amount of useful information provided by consistent information, namely:

$$\mathcal{I}(A^v \to A^u) = H(X_\Omega) - \mathcal{I}_D. \tag{56}$$

IV. CASE STUDY: THE BINARY SYMMETRIC CHANNEL

As an illustration of our general formalism, let us consider the standard example of a binary symmetric channel where we have two agents, A^v, A^u, sharing a world with two events, namely $\Omega = \{m_1, m_2\}$ such that $\mu(m_1) = \mu(m_2) = 1/2$.

Case 1: Non-preservation of referentiality. We will consider a case where $I(A^v \to A^u) = \max I$ but $\sigma_{A^v\to A^u} = \mathcal{I}(A^v \to A^u) = 0$. The transition matrices of agents A^v and A^u are identical and defined as:

$$A^{v,u} = \left\{\mathbf{P}^{v,u} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \mathbf{Q}^{v,u} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\right\}. \tag{57}$$

The channel between such agents, Λ, is noiseless:

$$\Lambda = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{58}$$

We begin by identifying the different elements involved in the process. First, from eq. (14) we obtain:

$$\mu'(m_1) = \mu'(m_2) = \frac{1}{2}.$$

The matrix of joint probabilities, $\mathbf{J}$, is -see eq. (33):

$$\mathbf{J} = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix}. \tag{59}$$

Thus, rearranging terms, the mutual information from A^u to A^v -see (eq. 28)- will be:

$$I(A^v \to A^u) = \log 2 = 1 \ \text{bit}.$$

We observe that, for a communication system consisting of two possible signals,

$$\max I = \log 2 = 1 \text{ bit}. \tag{60}$$

Thus the mutual information is maximum. However, it is evident that such a system does not preserve referentiality, since, if $X_\Omega = m_1$, then $X'_\Omega = m_2$, and viceversa. Indeed, let us first obtain the matrix $-\mathbf{J}\log\mathbf{J}$, which will be:

$$-\mathbf{J}\log\mathbf{J} = \begin{pmatrix} 0 & -\frac{1}{2}\log\frac{1}{2} \\ -\frac{1}{2}\log\frac{1}{2} & 0 \end{pmatrix}. \tag{61}$$

And, thus, by its definition, the referential term will be (eq. 34):

$$\sigma_{A^v\to A^u} = -\frac{\mathbf{tr}(\mathbf{J}\log\mathbf{J})}{\log 2} = 0, \tag{62}$$

(notice that $\log 2 = 1$, although we keep the logarithm for the sake of clarity) being the amount of consistent information:

$$\mathcal{I}(A^v \to A^u) = 0 \text{ bits}. \tag{63}$$

This extreme case dramatically illustrates the non-trivial relation between $\mathcal{I}$ and I, proposing a situation where the communication system is completely useless, although the mutual information between the random variables depicting the input and the output is maximum.

Case 2: Preservation of the referentiality. In this configuration, the referentiality is conserved. Let us suppose a different configuration of the agents. Now the transition matrices of agents A^v and A^u are identical and defined as:

$$A^{v,u} = \left\{\mathbf{P}^{v,u} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \mathbf{Q}^{v,u} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right\}. \tag{64}$$

The channel between such agents, Λ, is the two-dimensional noiseless channel defined in eq. (58). It is straightforward to check that the mutual information is maximal (= 1 bit), as above. The matrix $-\mathbf{J}\log\mathbf{J}$ will be, now,

$$-\mathbf{J}\log\mathbf{J} = \begin{pmatrix} -\frac{1}{2}\log\frac{1}{2} & 0 \\ 0 & -\frac{1}{2}\log\frac{1}{2} \end{pmatrix}. \tag{65}$$

This leads to $\sigma_{A^v\to A^u} = 1$, and, consequently:

$$I(A^v \to A^u) = \mathcal{I}(A^v \to A^u). \tag{66}$$

The above configuration is the only one which leads to $I = \mathcal{I}$. Furthermore -as shown in section III.b- it can only be achieved when I is maximum, i.e., in a noiseless scenario. In the last example we will deal with a noisy situation.

Case 3: Noisy channel. We finally explore the case where the matrix configuration of agents is the same as in the above example (eq. 57) but the channel is noisy, namely:

$$\Lambda = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}. \tag{67}$$

We first derive the matrix of joint probabilities, $\mathbf{J}$,which takes the following form:

$$\mathbf{J} = \begin{pmatrix} 0.45 & 0.05 \\ 0.05 & 0.45 \end{pmatrix}. \tag{68}$$

We now proceed by observing that $\mu'(m_1) = \mu'(m_2) = 1/2$. Thus, the mutual information will be:

$$\begin{aligned} I(A^v \to A^u) &= 0.9 \log \frac{0.45}{0.5 \cdot 0.5} + 0.10 \log \frac{0.05}{0.5 \cdot 0.5} \\ &= 0.531... \text{ bits.} \end{aligned} \tag{69}$$

To evaluate the degree of consistency of the communicative system, we firstly compute the matrix $-\mathbf{J} \log \mathbf{J}$:

$$\begin{pmatrix} -0.45 \log 0.45 & -0.05 \log 0.05 \\ -0.05 \log 0.05 & -0.45 \log 0.45 \end{pmatrix} = \begin{pmatrix} 0.518 & 0.216 \\ 0.216 & 0.518 \end{pmatrix}. \tag{70}$$

Since $H(X_\Omega, X'_\Omega) = 1.468$ bits, the referential parameter is:

$$\begin{aligned} \sigma_{A^v \to A^u} &= -\frac{\mathbf{tr}(\mathbf{J} \log \mathbf{J})}{H(X_\Omega, X'_\Omega)} \\ &= \frac{0.518 + 0.518}{1.468} \\ &= 0.706... \text{ consistent bits/bit.} \end{aligned} \tag{71}$$

(where the last “bit” refers to “bit obtained from the observation of input-output pairs”). The consistent information is, thus:

$$\begin{aligned} \mathcal{I}(A^v \to A^u) &= I(A^v \to A^u) \sigma_{A^v \to A^u} \\ &= 0.531 \times 0.706 \\ &= 0.375 \text{ bits.} \end{aligned} \tag{72}$$

Due to the symmetry of the problem, the average among the two agents is:

$$\mathcal{F}(A^v, A^u) = 0.375 \text{ bits.} \tag{73}$$

The amount of dissipated information is, thus:

$$\mathcal{I}_D = \overbrace{0.469}^{physical\ noise} + \overbrace{0.156}^{Referential\ noise} \text{ bits.} \tag{74}$$

We want to stress the following point: The matrix configuration is consistent with the framework proposed in *case 2*, where the amount of consistent information is maximum, but now the channel is noisy. The noisy channel has a double effect: first, it destroys information in the standard sense, since the noise parameter $H(X_\Omega | X'_\Omega) > 0$, but it also has an impact on the consistency of the process, introducing an amount of referential *noise* due to the lack of consistency derived from it. Thus, as derived in section III.b, eq. (40), in the presence of noise, we have shown that the inequalities

$$H(X_\Omega) > I(A^v \to A^u) > \mathcal{I}(A^v \to A^u) \tag{75}$$

hold, being, in our special case:

$$1 > 0.531 > 0.375. \tag{76}$$

V. DISCUSSION

The accurate definition of the amount of information carried by consistent input/output pairs is an important component of information transfer in biological or artificial communicating systems. In this paper we explore the central role of information exchanges in selective scenarios, highlighting the importance of the referential value of the communicative sign.

The conceptual novelty surrounding the paper can be easily understood from the role we attribute to *noise*. Physical information considers a source of $H(X)$ bits and a *dissipation* of $H(X|Y)$ bits due to, for example, thermal fluctuations. We add another source of information dissipation: the non-consistency of the pair signal/referent, putting aside the degree of correlation among random variables (see eq. 55). Indeed, in many physical processes no referentiality is at work, perhaps because, it is not relevant to wonder about the consistency of the communicative process. Moreover, if the whole system is *designed*, consistency problems are apriori ruled out, unless the engineer wants to explicitly introduce disturbances in the system. What makes biology different, however, is that biological systems are not designed but instead, are the outcomes of an evolutionary process where the nature of the response to a given stimulus is important, which makes the problem of consistency relevant for evolutionary scenarios. This problem needs an explicit formulation, being what we called *consistent information* the theoretical object that links raw information and function, or environmental response.

Are information processing mechanisms of living systems optimal regarding referentiality conservation? As we discussed above, it seems reasonable to assume that the conservation of referentiality must be at the core of any communicative system with some selective advantage. The general problem to find the optimal code, however, resembles the problem of finding the channel capacity, for which is well known that no general procedure exists (Cover and Thomas, 1991). Thus, how autonomous systems deal with such a huge mathematical

problem? One may consider the possibility that the co-evolution of the abstract coding and decoding entities; this would avoid the system to face a great amount of configurations per generation, thereby being all options highly limited at each generation where selection is at work.

We finally emphasize that the unavoidable dissipation of mutual information points to a reinterpretation of information-transfer phenomena in biological or self-organized systems, due to the important consequences that can be derived from it. Further work should explore the relevance of this limitation on more realistic scenarios, together with other implications that can be derived by placing equation (36) at the center of information transfer in biology.

Acknowledgments

We thank the members of the Complex Systems Lab for useful discussions. This work has been supported by a Juan de la Cierva grant from te Ministerio de Ciencia y Tecnología (JF), the James S. McDonnell Foundation (BCM) and by Santa Fe Institute (RS).

References

Ash, R. B., 1990, *Information Theory* (New York. Dover).

Cover, T. M., and J. A. Thomas, 1991, *Elements of Information Theory* (John Wiley and Sons. New York).

Haken, H., 1978, *Synergetics: An Introduction. Nonequilibrium Phase Transitions and Self- Organization in Physics, Chemistry and Biology (Springer Series in Synergetics)* (Springer).

Hopfield, J., 1994, Journal of Theoretical Biology **171**(1), 53,

Hurford, J., 1989, Lingua **77**(2), 187,

Kolmogorov, A., 1965, Problems Inform. Transmission **1**, 1.

Komarova, N. L., and P. Niyogi, 2004, Art. Int. **154**(1-2), 1,

Ming, L., and P. Vitányi, 1997, *An introduction to Kolmogorov complexity and its applications* (Springer, New York [u.a.]),

Niyogi, P., 2006, *The Computational Nature of Language Learning and Evolution* (MIT Press. Cambridge, Mass.),

Nolfi, S., and M. Mirolli, 2010, *Evolution of Communication and Language in Embodied Agents* (Berlin. Springer Verlag),

Nowak, M. A., 2000, Philosophical Transactions: Biological Sciences **355**(1403), 1615,

Nowak, M. A., and D. Krakauer, 1999, Proc. Nat. Acad. Sci. USA **96**(14), 8028,

Plotkin, J. B., and M. A. Nowak, 2000, Journal of Theoretical Biology **205**(1), 147,

Saussure, F., 1916, *Cours de Linguistique Générale* (Bibliothèque scientifique Payot: Paris),

Shannon, C. E., 1948, Bell System Technical Journal **27**, 379.

Steels, L., 2001, IEEE Intelligent Systems **16**, 16, ISSN 1541-1672.

Steels, L., and J.-C. Baillie, 2003, Robotics and Autonomous Systems **43**(2-3), 163.

Von Neumann, J., 1936, *The Mathematical Foundations of Quantum Mechanics* (Princeton University Press. Princeton, N. J.).

Yockey, H. P., 1992, *Information Thery and Molecular Biology* (Cambridge University Press, Cambridge (UK)).