PROCEEDINGS
OF
THE ROYAL
SOCIETY B

Proc. R. Soc. B (2005) **272**, 561–565
doi:10.1098/rspb.2004.2957
Published online 7 March 2005

The consequences of Zipf's law for syntax and symbolic reference

Ramon Ferrer i Cancho[1][2]*, Oliver Riordan[3] and Béla Bollobás[3,4]

[1] *ICREA-Complex Systems Laboratory, IMIM-UPF, Dr Aiguader 80, 08003 Barcelona, Spain*
[2] *INFM udR Roma1, Dipartimento di Fisica, Università "La Sapienza", Piazzale Aldo Moro 5, 00185 Roma, Italy*
[3] *Department of Pure Mathematics and Mathematical Statistics, Cambridge and Trinity College, Cambridge CB2 1TQ, UK*
[4] *Department of Mathematical Sciences, University of Memphis, Memphis, TN 38152, USA*

Although many species possess rudimentary communication systems, humans seem to be unique with regard to making use of syntax and symbolic reference. Recent approaches to the evolution of language formalize why syntax is selectively advantageous compared with isolated signal communication systems, but do not explain how signals naturally combine. Even more recent work has shown that if a communication system maximizes communicative efficiency while minimizing the cost of communication, or if a communication system constrains ambiguity in a non-trivial way while a certain entropy is maximized, signal frequencies will be distributed according to Zipf's law. Here we show that such communication principles give rise not only to signals that have many traits in common with the linking words in real human languages, but also to a rudimentary sort of syntax and symbolic reference.

Keywords: Zipf's law; syntax; symbolic reference; human language

1. INTRODUCTION

Word frequencies in human languages tend to obey Zipf's law (Zipf 1972), which states that for some $\beta > 0$,

$$p_f \sim f^{-\beta}, \tag{1.1}$$

where p_f is the proportion of words whose frequency in a given sample is f. Usually $\beta \approx 2$ is found (Ferrer i Cancho 2005).

Zipf's law has been shown to appear when simultaneously maximizing the communicative efficiency and minimizing the cost of communication (Ferrer i Cancho & Solé 2003). Alternatively, constraining the ambiguity of communication in a non-trivial way while a certain entropy is maximized will also lead to Zipf's law, with a wider range of exponents (Ferrer i Cancho 2005). With Zipf's law (or the communication principles leading to that law) as the basic assumption, we explore its consequences for a simple communication system.

Our aim is to show that a basic assumption (a form of Zipf's law) naturally leads to certain consequences, in particular, a certain combinatorial property of words, connectedness, that is a precondition for syntax; this is described in detail below. To this end, we shall study a highly simplified and abstracted linguistic model: both the assumption and the consequences make sense in this setting, without reference to the much more complicated details of real languages, or more realistic models for them. Our model will not, of course, be strictly realistic for any particular language, or even for the early developing human language to which we think our conclusions are most relevant and about which very little is known. In some sense, we wish to show that connectedness arises naturally from Zipf's law, independently of the details of the linguistic setting. To do this, we shall consider a random model, within a class specified below. Once again, no real language is formed in this random way, but this is the point. As almost any model of the given form shows connectedness, the absence of connectedness would need further explanation, but given Zipf's law, connectedness does not, under a wide range of conditions.

2. THE MODEL

We assume a general communication framework and thus define a set of signals, $S = \{s_1, \ldots, s_i, \ldots, s_n\}$, and a set of objects, $R = \{r_1, \ldots, r_j, \ldots, r_m\}$. A signal is a generic code that is capable of carrying meaning. The general term signal is used here to provide a high enough level of abstraction. For instance, we want to abstract from the signal medium (vocal, gestural, chemical) or the type of reference involved (iconic, indexical or symbolic; Deacon 1997). In order to exclude codes like syllables or sentences in human language from the kind of signals intended here, our signals should not be decomposable into simpler units unless such units do not have referential power. Human words can only be replaced by signals in a metaphorical sense because human words imply symbolic reference, whereas our signals are not necessarily symbols. Objects here may be cognitive categories (Damper & Harnad 2000; Harnad 2003) and therefore be modelled by a discrete set.

We define a matrix of signal–object associations $A = \{a_{ij}\}$ $(1 \le i \le n,\ 1 \le j \le m)$ where $a_{ij} = 1$ if the i-th signal and the j-th object are associated (the j-th object is a 'possible meaning' for the i-th signal) and $a_{ij} = 0$ otherwise. We consider 'binary' associations only for simplicity. The matrix A defines a bipartite graph $G_{n,m}$ (Bollobás 1998) with edges corresponding to the ones in A. This matrix A defines signal–object associations that can be of

*Author and address for correspondence: INFM udR Roma1, Dipartimento di Fisica, Universitá "La Sapienza", Piazzale Aldo Moro 5, 00185 Roma, Italy (ramon@pil.phys.uniromal.it).

Received 18 May 2004
Accepted 28 September 2004

two types: referential and non-referential. By referential we mean that the signal can refer to the object, as in the link between the word 'meat' and the object 'edible organic matter', or as in the link between the verb 'eat' and the object 'the action of eating'. By non-referential we mean the remaining possible signal–object associations. For instance, the syntactic association between a verb and its argument would be realized in our model by signal–object associations between the verb and the objects representing the possible arguments. Thus, the verb 'eat' is associated not only to the object the 'action of eating' (referentially) but also to the object 'edible organic matter' (non-referentially).

Various theoretical approaches to syntax assume that a connection between a pair of syntactically linked words implies that the words are semantically compatible (Chomsky 1965; Helbig 1992). Here we assume that a connected pair of signals are connected to each other through a common object, which, acting as a rudimentary meaning, defines the semantic compatibility of the pair. Therefore, signals having a common object may or may not be synonyms (depending on whether the pair of links is referential or not). We can also model forbidden arguments. For instance, the object 'umbrella' cannot be the object of the verb 'eat', so there would be no link between this object and the verb 'eat'. In a very simplified manner, $G_{n,m}$ contains information about argument structure.

Objects are simple meanings. Words in human language have complex meanings that may involve more than one of our objects here. For instance, the word 'eat' in human language is associated with at least two objects in our view: 'the action of eating' (referentially), and 'edible organic matter' (non-referentially). Both objects are needed to understand the meaning of the verb 'eat'.

The matrix A should be seen as a primitive association system from which different types of signal–object associations may develop. Note that without the skeleton provided by A, complex types of associations (synonymy links, syntactic links) cannot emerge. A, in spite of being a simplification, may be arranged in a way that may lead to a simple form of language or in a way that cannot. Here we study how and why Zipf's law leads to the former case.

Let us write p_k for the proportion of signals with k links. We make the natural simplifying assumption that the relative frequency of a signal is proportional to the number of objects it is connected to, as in Ferrer i Cancho (2005). Under this assumption, Zipf's law (equation (1.1)) is equivalent to

$$p_k \sim k^{-\beta}. \tag{2.1}$$

In what follows, we shall assume equation (2.1). Our model for $G_{n,m}$ will be as follows: given the numbers n and m of signals and objects, and for each k, the proportion p_k of signals connected to k objects, the graph $G_{n,m}$ is chosen uniformly at random from among all bipartite graphs with these properties. Equivalently, having decided the degree, $d(s_i)$ (i.e. the number of associated objects), of each signal appropriately, we join s_i to a random set of $d(s_i)$ objects, independently of the other signals. We investigate properties that the resulting graph has with high probability, noting that any such property is a very natural consequence of Zipf's law. Indeed, as noted in the introduction, the model is not complete or strictly realistic, and one cannot deduce that such a property has developed in the real world only because of Zipf's law. However, Zipf's law is a sufficient explanation: given Zipf's law, it would be more surprising if the property did not hold than if it did.

Note that there is a transition in the model at $\beta=2$, owing to the rapid change in the number of edges as β is varied about this value. More precisely, the average degree of a signal is $\sum_{k=1}^{m} kp_k$. The infinite form of this sum converges if, and only if, $\beta>2$; in this range the average degree is asymptotically constant as m increases. In contrast, for $\beta=2$ the average degree grows logarithmically with m and, for $\beta<2$, as a power of m. In asymptotic analysis we shall thus consider $\beta=2+\varepsilon$ for some small ε.

Given the signal–object graph $G_{n,m}$, we define a signal–signal graph G_n whose vertices are the signals s_i, in which two signals are joined if in $G_{n,m}$ they are joined to one or more common objects. Some links in G_n are synonymy links because they stem from two referential links in $G_{n,m}$, but the remaining links define syntactic links (e.g. verb–argument links). We define G_n' as the subgraph of G_n formed by all signals in S and all the non-synonymy links in G_n. What follows is true if the proportion of synonymy links in G_n is small.

In our simplified model, a grammatical phrase can be formed by choosing a pair of signals (u, v) in G_n' and all the signals in a path from u to v. Total freedom for forming phrases only exists when there is a path between every pair of vertices, that is, when the network is connected. Connectedness is a stronger and more realistic precondition for syntax than just combinations of a few signals, as in recent formal approaches to the origins of syntax (Nowak & Krakauer 1999; Nowak *et al.* 2000). For various reasons, our grammar is not a grammar in the strict sense, but rather a protogrammar, from which full human language can easily evolve. First, note that such a grammar lacks word order (Sleator & Temperley 1991) and link direction (Melčuk 1989). Second, such a grammar does not imply (but allows) recursion (Hauser *et al.* 2002). Handling recursion implies memory resources (Lieberman 1991) that are not necessarily available when connectedness is reached.

When $\beta=2+\varepsilon$, there is a high probability that G_n' is almost connected, in the sense that almost all signals lie in a single component (the limiting proportion tends to one as $\varepsilon\to 0$; see figure 1*a*,*b*). Almost connectedness is easy to derive mathematically, although there is no space here for the details. We shall work in G_n rather than G_n' for simplicity. All our results carry over to G_n' if the proportion of synonymy links is small; the asymptotic results (such as that just stated) carry over for any fixed proportion of synonymy links less than one.

There are two key requirements for almost connectedness. Firstly, the 'expected neighbourhood expansion factor' f must be greater than one. Roughly speaking, f is the average number of nodes (here signals) within distance $\ell+1$ of a given node s, divided by the number within distance ℓ, for ℓ in a suitable range. If ℓ is neither too small nor too large and t is a node at distance ℓ from s, then the expected number of neighbours of t at distance $\ell+1$ from s is essentially independent of ℓ. Here, noting that one 'step'

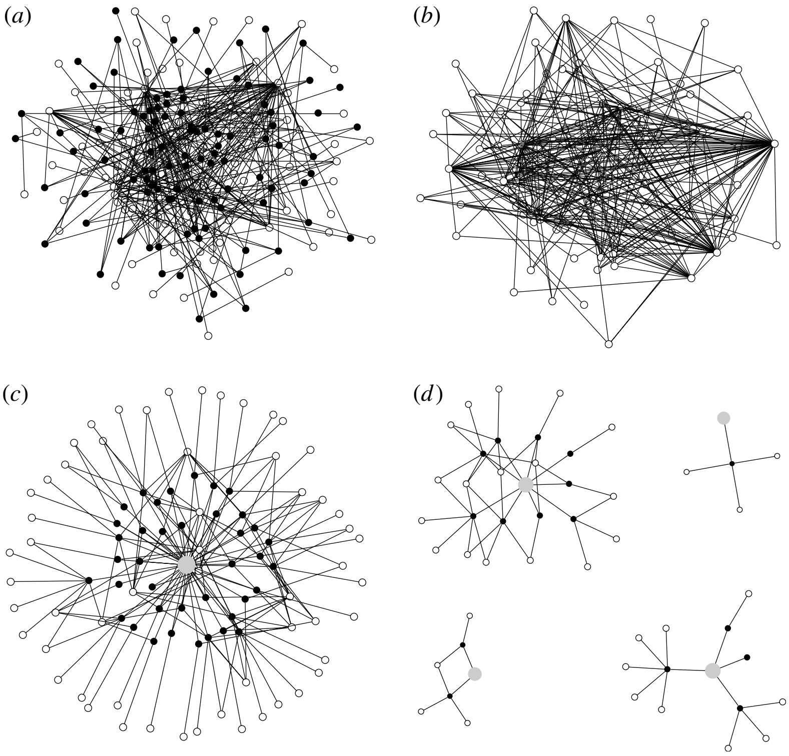

Figure 1. (*a*) Examples of $G_{n,m}$ and (*b*) G_n for $\beta=2$ and $n=m=100$. White and black circles are signals and objects, respectively. (*c*) First and second neighbours of the most connected signal (grey circle) in *a*. This and other highly connected signals are the forerunners of linking words (e.g. prepositions and conjunctions) in human language. (*d*) First and second neighbours of other signals (grey circles) in *a*.

in G_n corresponds to two in $G_{n,m}$, one can check that

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

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

Connectedness or near connectedness also implies a higher order association where signal–signal associations emerge from signal–object associations. If s_i and r_j are linked, and r_j and s_k are also linked ($i \neq k$), then there is a signal–signal association between s_i and s_k formed via r_j in only two steps. Signal–signal associations are the basis of a rudimentary form of symbolic reference. Symbolic reference is about how a word not only evokes a certain 'meaning', but also how that word evokes other words (Deacon 1997).

We will show that the degree distribution in syntactic dependency networks (Ferrer i Cancho *et al.* 2003) easily follows from assuming equation (2.1). In short, these networks are formed by words as vertices, and two words are linked if they have been syntactically combined in a collection of sentences. The links in the sentence 'John eats apples' consist of two syntactic dependency links, one between 'John' and 'eats' and another between 'eats' and 'apples' (the former between the subject of the sentence and its verb and the latter between the verb and its object; see Melčuk (1989) for a description of the syntactic dependency formalism). These links would belong to the syntactic dependency network if the sentence were one of the sentences in the collection.

We define q_k as the proportion of signals having degree k in G_n, recalling that two signals are joined in G_n if they are associated with at least one common object in $G_{n,m}$. Let Z be the degree in G_n of a random signal s_i, so $q_k = \Pr(Z=k)$. With $\beta=2+\varepsilon$ it is very unlikely that two given signals are joined to two or more common objects, so Z is essentially

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

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

$$q_k \sim ck^{-\beta}, \tag{2.2}$$

with c a positive constant. Thus, while the exact distribution of Z is not a power law, Z does have a power-law tail, with the same exponent β as the signal degrees and the signal frequency distribution (equation (1.1)). Equation (2.2) is consistent with the analysis of real syntactic dependency networks, where the proportion of words having k syntactic links with other words is $\sim k^{-\gamma}$ with $\gamma \approx 2.2$ (Ferrer i Cancho *et al.* 2003). Note that γ is in turn close to the typical Zipf's law exponent.

Syntactic theory regards certain function words, such as prepositions and conjunctions, as linkers (Melčuk 1989); that is, words serving as 'combining words' for forming complex sentences. The most connected signals in $G_{n,m}$ share many features with real linking words (figure 1). Linkers in human language have (a) poor (or absent) referential power (Givón 2002), (b) high frequency (Baayen 2001), and (c) many connections with referentially powerful words. (a) and (b) are satisfied by the most connected words in $G_{n,m}$ based only on two basic axioms: (i) Zipf's law in the distribution of the number of connections per signal in $G_{n,m}$, and (ii) a proportionality relationship between signal frequency of use and number of connections in $G_{n,m}$. (c) requires a further axiom: (iii) two vertices in G'_n are linked if they have at least one common object in $G_{n,m}$. High degree vertices in G'_n satisfy (a) since the uncertainty associated with the interpretation of a signal grows with its number of links in $G_{n,m}$ (Ferrer i Cancho 2005). The most connected links in G_n are also the most connected links in G'_n. Satisfying (b) follows trivially from (i) and (ii). (c) follows from the skewed and heavy-tailed distributions for q_k, which is in turn a consequence of (i).

3. DISCUSSION

We have seen that, in our simplified model, Zipf's law is a sufficient condition for almost connectedness provided the number of signals and objects are similar, and that Zipf's law with almost connectedness implies the existence of linking words. Almost connectedness in signal–object associations is a necessary precondition for full syntax and for going beyond mere simple signal–object associations.

The two-level organization of linguistic structure, with a limited set of words created by combinations of meaningless syllables at one level, and a limitless set of sentences created by combining words at the other, is a critical feature of language, sometimes termed 'duality of patterning' (Hockett 1960). Duality of patterning has not been fully considered as requisite in recent models of the origins of syntax (Nowak & Krakauer 1999; Nowak *et al.* 1999, 2000; Nowak 2000). Here, the essential requirement of connectedness is consistent with duality of patterning. We have defined phrases as paths in G'_n. Different paths in G'_n correspond to syntactically different phrases, whereas different paths in $G_{n,m}$ starting and ending at signals correspond to semantically different phrases. Note that a certain path in G'_n corresponds to at least one path in $G_{n,m}$, so, assuming the proportion of synonymy links is small, the expressivity is given by the number of signal–signal paths in $G_{n,m}$. When approximating semantically different phrases by signal–signal paths in $G_{n,m}$, the number of phrases formed by paths allowed to pass more than once through the same vertex is, of course, infinite. Since paths repeating vertices are to some extent redundant, perhaps the more interesting case is that of paths passing at most once through each vertex. In this case, it is easy to show that, whenever there is a giant component, there is a constant $c>1$ such that the expected number of these paths is at least c^n. To summarize, although our model is obviously much simpler than present-day languages, it provides a basis for the astronomically large number of sentences that human speakers can produce and process.

While researchers are divided when considering syntax (Nowak & Krakauer 1999; Nowak *et al.* 2000; Hauser *et al.* 2002) or symbolic reference (Donald 1991, 1998; Deacon 1997) as the essence of human language, we hypothesize that syntax and forms of reference higher than mere signal–object associations are two sides of the same coin, i.e. connectedness in signal–signal associations. A communication system maximizing the information transfer (i.e. minimizing the effort for the hearer) by mapping every object to a distinctive signal (Ferrer i Cancho & Solé 2003) implies that two signals in $G_{n,m}$ never share the same object, so G_n (and G'_n) have no links at all. Therefore, a perfect communication system cannot be connected, or even almost connected in any sense. Such a system cannot satisfy the simple precondition for syntax and complex reference that Zipf's law provides. Many non-human species seem to be close to a perfect communication system for two reasons. One is practical: those species have difficulties in dealing with signal ambiguity (Deacon 1997). The other one is theoretical: when minimizing hearer and speaker needs simultaneously, there seem to be only two possible basic configurations: no communication and perfect communication. Zipf's law (with non-extremal exponents; Ferrer i Cancho 2005), and therefore human language, appears in a very narrow domain between these two configurations, so non-human communication is more likely to be in the perfect communication phase than in the narrow Zipfian domain (Ferrer i Cancho & Solé 2003). As far as we know, no non-human species arranges its meaningful signals according to equation (1.1) with $\beta = 2$.

Zipf's law provides connectedness, an essential precondition for syntax and complex reference, for free. Hence, as language developed, the transition to syntax and complex types of reference may perhaps have been as abrupt as the transition to Zipf's law (Ferrer i Cancho & Solé 2003). While some researchers consider Zipf's law a meaningless pattern in human language (Mandelbrot 1953; Miller & Chomsky 1963; Li 1992; Nowak *et al.* 2000), we have shown that Zipf's law provides a simple communication system with fundamental traits that do not arise in perfect communication systems.

The authors are grateful to Toni Hernández for helpful comments. This work was supported by grants of the Generalitat de Catalunya (FI/2000-00393; R.F.C.), the FET Open Project COSIN (IST-2001-33555; R.F.C.), DARPA grant F33615-01-C-1900 (B.B.), NSF grant ITR 0225610 (B.B.), and a Research Fellowship of the Royal Society (O.R.).

REFERENCES

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

Bollobás, B. 1998 *Modern graph theory*. New York: Springer.

Chomsky, N. 1965 *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Damper, R. I. & Harnad, S. R. 2000 Neural network modeling of categorical perception. *Percept. Psychophys.* **62**, 843–867.

Deacon, T. W. 1997 *The symbolic species: the co-evolution of language and the brain*. New York: W.W. Norton & Company.

Donald, M. 1991 *Origins of the modern mind*. Cambridge, MA: Harvard University Press.

Donald, M. 1998 Mimesis and the executive suit: missing links in language evolution. In *Approaches to the evolution of language: social and cognitive bases* (ed. J. R. Hurford, M. Studdert-Kennedy & C. Knight), pp. 44–67. Cambridge University Press.

Ferrer i Cancho, R. 2005 Decoding least effort and scaling in signal frequency distributions. *Physica A*. **345**, 275–284.

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

Ferrer i Cancho, R., Solé, R. V. & Köhler, R. 2003 Patterns in syntactic dependency networks. *Phys. Rev. E* **69**, 051915-1–051915-8.

Givón, T. 2002 *Bio-linguistics*. Amsterdam: John Benjamins.

Harnad, S. 2003 *Categorical perception Encyclopedia of cognitive science*. London: Nature Publishing Group/Macmillan.

Hauser, M. D., Chomsky, N. & Fitch, W. T. 2002 The faculty of language: what is it, who has it and how did it evolve? *Science* **298**, 1569–1579.

Helbig, G. 1992 *Probleme der valenz und kasustheorie*. Tübinguen: Niemeyer.

Hockett, C. F. 1960 Logical considerations in the study of animal communication. In *Animal sounds and communications* (ed. W. Lanyon & W. Tavolga), pp. 392–430. Washington: American Institute of Biological Sciences.

Li, W. 1992 Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE T. Inform. Theory* **38**(6), 1842–1845.

Lieberman, P. 1991 *Uniquely human: the evolution of speech, thought and selfless behavior*. Cambridge, MA: Harvard University Press.

Mandelbrot, B. 1953 An informational theory of the statistical structure of language. In *Communication theory* (ed. W. Jackson), p. 486. London: Butterworths.

Melčuk, I. 1989 *Dependency grammar: theory and practice*. University of New York Press.

Miller, G. A. & Chomsky, N. 1963 Finitary models of language users. In *Handbook of mathematical psychology* (ed. R. D. Luce, R. Bush & E. Galanter), vol. 2. New York: Wiley.

Nowak, M. A. 2000 Evolutionary biology of language. *Phil. Trans. R. Soc. B* **355**, 1615–1622.

Nowak, M. A. & Krakauer, D. C. 1999 The evolution of language. *Proc. Natl Acad. Sci. USA* **96**, 8028–8033.

Nowak, M. A., Krakauer, D. C. & Dress, A. 1999 An error limit for the evolution of language. *Proc. R. Soc. B* **266**, 2131–2136.

Nowak, M. A., Plotkin, J. B. & Jansen, V. A. 2000 The evolution of syntactic communication. *Nature* **404**, 495–498.

Sleator, D. & Temperley, D. 1991 *Parsing English with a link grammar* (Tech. Rep.). Carnegie Mellon University.

Zipf, G. K. 1972 *Human behaviour and the principle of least effort. An introduction to human ecology*. New York: Hafner (1st edition; Reprinted by Cambridge, MA: Addison-Wesley 1949).