Least effort and the origins of scaling in human language

Ramon Ferrer i Cancho[†‡] and Ricard V. Solé[†§]

[†]Complex Systems Lab, Universitat Pompeu Fabra, Doctor Aiguader 80, 08003 Barcelona, Spain; and [§]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved December 6, 2002 (received for review September 29, 2002)

The emergence of a complex language is one of the fundamental events of human evolution, and several remarkable features suggest the presence of fundamental principles of organization. These principles seem to be common to all languages. The best known is the so-called Zipf's law, which states that the frequency of a word decays as a (universal) power law of its rank. The possible origins of this law have been controversial, and its meaningfulness is still an open question. In this article, the early hypothesis of Zipf of a principle of least effort for explaining the law is shown to be sound. Simultaneous minimization in the effort of both hearer and speaker is formalized with a simple optimization process operating on a binary matrix of signal–object associations. Zipf's law is found in the transition between referentially useless systems and indexical reference systems. Our finding strongly suggests that Zipf's law is a hallmark of symbolic reference and not a meaningless feature. The implications for the evolution of language are discussed. We explain how language evolution can take advantage of a communicative phase transition.

Beyond their specific differences, all known human languages exhibit two fully developed distinguishing traits with regard to animal communication systems: syntax (1) and symbolic reference (2). Trying to explain the complexity gap between humans and other species, different authors have adopted different views from gradual evolution (3) to non-Darwinian positions (4). Arguments are often qualitative in nature and sometimes ad hoc. Only recently mathematical models have explicitly addressed these questions (5, 6).

It seems reasonable to assume that our human ancestors started off with a communication system capable of rudimentary referential signaling, which subsequently evolved into a system with a massive lexicon supported by a recursive system that could combine entries in the lexicon into an infinite variety of meaningful utterances (7). In contrast, nonhuman repertoires of signals are generally small (8, 9). We aim to provide new theoretical insights to the absence of intermediate stages between animal communication and language (9).

Here we adopt the view that the design features of a communication system are the result of interaction between the constraints of the system and demands of the job required (7). More precisely, we will understand the demands of a task such as providing easy-to-decode messages for the receiver. Our system will be constrained by the limitations of a sender trying to code such an easy-to-decode message.

Many authors have pointed out that tradeoffs of utility concerning hearer and speaker needs to appear at many levels. As for the phonological level, speakers want to minimize articulatory effort and hence encourage brevity and phonological reduction. Hearers want to minimize the effort of understanding and hence desire explicitness and clarity (3, 10). Regarding the lexical level (10, 11), the effort for the hearer has to do with determining what the word actually means. The higher the ambiguity (i.e., the number of meanings) of a word, the higher the effort for the hearer. Besides, the speaker will tend to choose the most frequent words. The availability of a word is positively correlated with its frequency. The phenomenon known as the *word-frequency effect* (12) supports it. The most frequent words tend to be the most ambiguous ones (13). Thereafter, the speaker tends to choose the most ambiguous words, which is opposed to the least effort for the hearer. Zipf referred to the lexical tradeoff as the *principle of least effort*. He pointed out that it could explain the pattern of word frequencies, but he did not give a rigorous proof of its validity (11). Word frequencies obey Zipf's law. If the words of a sample text are ordered by decreasing frequency, the frequency of the kth word, $P(k)$, is given by $P(k) \propto k^{-\alpha}$, with $\alpha \approx 1$ (11). This pattern is robust and widespread (14).

Here we show that such a lexical compromise can be made explicit in a simple form of language game where minimization of speaker and hearer needs is introduced in an explicit fashion. As a consequence of this process and once a given threshold is reached, Zipf's law, a hallmark of human language, emerges spontaneously.

The Model

To define explicitly the compromise between speaker and hearer needs, a cost function must be introduced. Given the nature of our systems, information theory provides the adequate mathematical framework (15). We consider a system involving a set of n signals $\mathcal{S} = \{s_1,\ldots, s_i,\ldots, s_n\}$ and a set of m objects of reference $\mathcal{R} = \{r_1,\ldots, r_i,\ldots, r_m\}$. The interactions between signals and objects of reference (hereafter objects) can be modeled with a binary matrix $\mathbf{A} = \{a_{ij}\}$, where $1 \leq i \leq n$ and $1 \leq j \leq m$. If $a_{ij} = 1$, then the ith signal refers to the jth object, and $a_{ij} = 0$ otherwise. We define $p(s_i)$ and $p(r_j)$ as the probability of s_i and r_j, respectively. If synonymy were forbidden, we would have

$$p(s_i) = \sum_j a_{ij} p(r_j), \qquad [1]$$

because signals are used for referring to objects. We assume $p(r_i) = 1/m$ in what follows. If synonymy is allowed, the frequency of an object has to be distributed among all its signals. The frequency of a signal, $p(s_i)$ is defined as

$$p(s_i) = \sum_j p(s_i, r_j). \qquad [2]$$

According to the Bayes theorem we have

$$p(r_j, s_i) = p(r_j) p(s_i \mid r_j). \qquad [3]$$

$p(s_i \mid r_j)$ is defined as

$$p(s_i \mid r_j) = a_{ij} \frac{1}{\omega_j}, \qquad [4]$$

where $\omega_i = \Sigma_j a_{ji}$ is the number of synonyms of j. Substituting Eq. **4** into Eq. **3** we get

This paper was submitted directly (Track II) to the PNAS office.

[‡]To whom correspondence should be addressed. E-mail: ramon.ferrer@cexs.upf.es.

www.pnas.org/cgi/doi/10.1073/pnas.0335980100

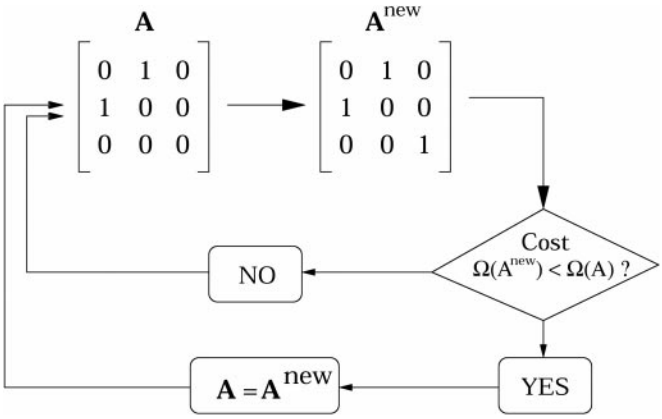

Fig. 1. Basic scheme of the evolutionary algorithm used in this article. Starting from a given signal–object matrix **A** (here $n = m = 3$), the algorithm performs a change in a small number of bits (specifically, with probability ν, each a_{ij} can flip). The cost function Ω is then evaluated, and the new matrix is accepted provided that a lower cost is achieved. Otherwise, we start again with the original matrix. At the beginning, **A** is set up with a fixed density ρ of ones.

$$p(r_j, s_i) = a_{ij}\frac{p(r_j)}{\omega_j}. \quad [5]$$

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

$$H_n(S) = -\sum_{i=1}^{n} p(s_i)\log_n p(s_i). \quad [6]$$

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

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

$$H_m(\mathcal{R} \mid s_i) = -\sum_{j=1}^{m} p(r_j \mid s_i)\log_m p(r_j \mid s_i), \quad [7]$$

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

$$H_m(\mathcal{R} \mid \mathcal{S}) = \sum_{i=1}^{n} p(s_i)H_m(\mathcal{R}, s_i). \quad [8]$$

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

$$\Omega(\lambda) = \lambda H_m(\mathcal{R} \mid \mathcal{S}) + (1 - \lambda)H_n(\mathcal{S}), \quad [9]$$

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

Methods

$\Omega(\lambda)$ is minimized with the following algorithm, summarized in Fig. 1. At each step, the graph is modified by randomly changing the state of some pairs of vertices, and the new **A** matrix is accepted if the cost is lowered [if an object has no signals, $\Omega(\lambda) = \infty$]. The algorithm stops when the modifications on **A** are not accepted $T = 2nm$ times in a row. Configurations for which an object has no signals assigned are forbidden.

If Zipf's hypothesis were valid, a Zipfian distribution of signal frequencies should appear for $\lambda \approx 1/2$, where the efforts for the speaker and the hearer have a similar contribution to the cost function. Notice that $\Omega(1/2) = H_{n \bullet m}(\mathcal{S}, \mathcal{R})/2$.

Results

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

$$I_n(\mathcal{S}, \mathcal{R}) = H_n(\mathcal{S}) - H_n(\mathcal{S} \mid \mathcal{R}), \quad [10]$$

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

$$L = \frac{|\{i \mid \mu_i > 0\}|}{n} \quad [11]$$

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

Three domains can be distinguished in the behavior of $I_n(\mathcal{S}, \mathcal{R})$ versus λ, as shown in Fig. 2*A*. First, $I_n(\mathcal{S}, \mathcal{R})$ grows smoothly for $\lambda < \lambda^* \approx 0.41$. $I_n(\mathcal{S}, \mathcal{R})$ explodes abruptly for $\lambda = \lambda^* \approx 0.41$. An abrupt change in L (Fig. 2*A*) versus λ (Fig. 2*B*) is also found for $\lambda = \lambda^*$. Single-signal systems ($L \approx 1/n$) dominate for $\lambda < \lambda^*$. Because every object has at least one signal, one signal stands for all the objects. $I_n(\mathcal{S}, \mathcal{R})$ indicates that the system is unable to convey information in this domain. Rich vocabularies ($L \approx 1$) are found for $\lambda > \lambda^*$. Full vocabularies are attained beyond $\lambda \approx 0.72$. The maximal value of $I_n(\mathcal{S}, \mathcal{R})$ indicates that the associations between signals and objects are one-to-one maps.

As for the signal frequency distribution in every domain, very few signals have nonzero frequency for $\lambda < \lambda^*$ (Fig. 3*A*), scaling consistent with Zipf's law appears for $\lambda = \lambda^*$ (Fig. 3*B*), and an almost uniform distribution is obtained for $\lambda > \lambda^*$ (Fig. 3*C*). As it occurs with other complex systems (16), the presence of a phase transition is associated with the emergence of power laws (17).

Knowing that $I_n(\mathcal{S}, \mathcal{R}) = I_n(\mathcal{R}, \mathcal{S})$ and using Eq. **10**, minimizing Eq. **9** is equivalent to minimizing

$$\Omega(\lambda) = \lambda I_n(\mathcal{S}, \mathcal{R}) + (1 - \lambda)H_n(\mathcal{S}). \quad [12]$$

Other functions could be proposed. Interestingly, the symmetric version of Eq. **9** with conditional entropies in both terms of the right side,

$$\Omega(\lambda) = \lambda H_m(\mathcal{R} \mid \mathcal{S}) + (1 - \lambda)H_n(\mathcal{S} \mid \mathcal{R}), \quad [13]$$

will help us to understand the origins of the sharp transition. Although the global minimum of $H_n(\mathcal{S})$ (one signal for all objects) is a maximum of $H_m(\mathcal{R} \mid \mathcal{S})$, the global minimum of $H_m(\mathcal{R} \mid \mathcal{S})$ (signal–object one-to-one maps with $n = m$) is a maximum of $H_n(\mathcal{S})$ in Eq. **9**. Thus both terms of Eq. **9** are in conflict. In contrast, the global minimum of $H_n(\mathcal{S} \mid \mathcal{R})$ is a subset of the global minimum of $H_m(\mathcal{R} \mid \mathcal{S})$ in Eq. **13**. Consistently, numerical optimization of Eq. **13** shows no evidence of scaling for Eq. **13**. Not surprisingly, the minimization of Eq. **13** is equivalent to

$$\Omega(\lambda) = I_n(\mathcal{S}, \mathcal{R}) + (1 - \lambda)H_n(\mathcal{S}). \quad [14]$$

Notice that λ is present in only one of the terms of the right side of the previous equation. Zipf's hypothesis was based on a tension between unification and diversification forces (11) that Eq. **13** does not accomplish. Eq. **9** does.

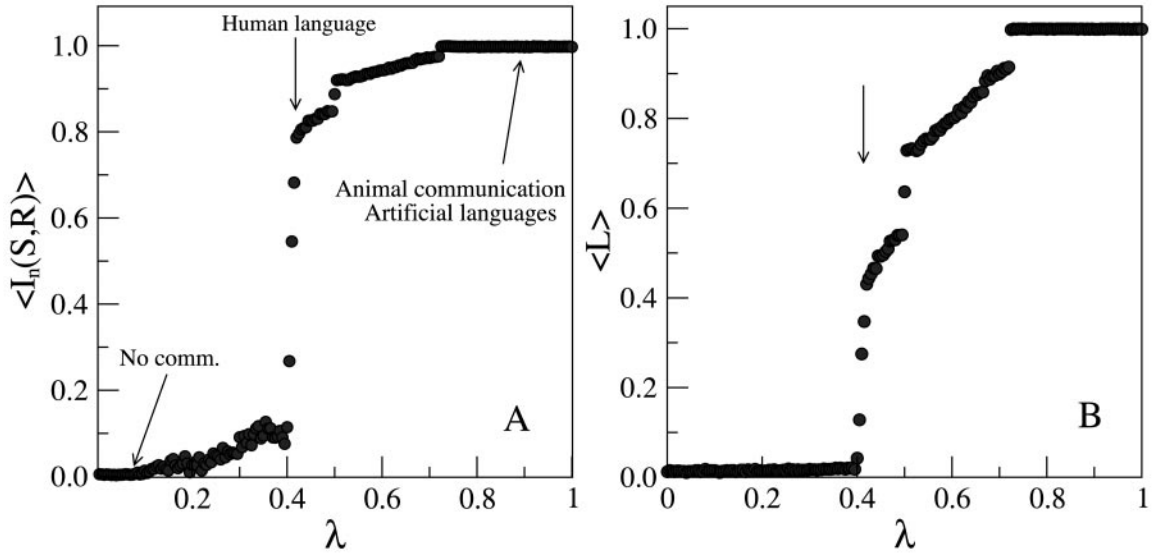

Fig. 2. (*A*) $\langle I_n(S, R)\rangle$, the average mutual information as a function of λ. $\lambda^* = 0.41$ divides $\langle I_n(S, R)\rangle$ into no-communication and perfect-communication phases. (*B*) Average (effective) lexicon size, $\langle L\rangle$, as a function of λ. An abrupt change is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

Discussion

Theoretical models support the emergence of complex language as the result of overcoming error limits (5) or thresholds in the amount of objects of reference that can be handled (6). Despite their power, these models make little use of some well known quantitative regularities displayed by most human languages such as Zipf's law (11, 18). Most authors, however, make use of Zipf's law as a null hypothesis with no particular significance (6). As far as we know, there is no compelling explanation for Zipf's law, although many have been proposed (19–23). Random texts (random combinations of letters and blanks) reproduce Zipf's law (19, 24–26) and are generally regarded as a null hypothesis (18). Although random texts and real texts differ in many aspects (26, 27), the possibility that Zipf's law results from a simple process (not necessarily a random text) has not been soundly denied. Our results show that Zipf's law is the outcome of the nontrivial arrangement of word–concept associations adopted for complying with hearer and speaker needs. Sudden changes in Fig. 2 and the presence of scaling (Fig. 3*B*) strongly suggest that a phase transition is taking place at $\lambda = \lambda^*$ (17).

Maximal mutual information (that is, one-to-one signal–object maps) beyond the transition is the general outcome of artificial-life language models (28, 29) and the case of animal communication (2), where small repertoires of signals are found (8, 9). On the one hand, speaker constraints ($\lambda < \lambda^*$) are likely to cause species with a powerful articulatory system (providing them with a big potential vocabulary) to have a referentially useless communication system (8). On the other hand ($\lambda > \lambda^*$), least effort for the hearer forces a species to have a different signal for each object at the maximum effort at the expense of the speaker, which allows us to make the following predictions. First, nonhuman repertoires must be small to cope with maximum speaker costs. Consistently, their size is on the order of 20–30 signals for the larger repertoires (8). Second, the large lexicons used by humans cannot be one-to-one maps because of the word-frequency effect (12) that makes evident how lexical access-retrieval cost is at play in humans. Third, large lexicons with one-to-one maps can be obtained only under idealized conditions when effort for the speaker is neglected. This is the case of artificial-language communication models, which reach

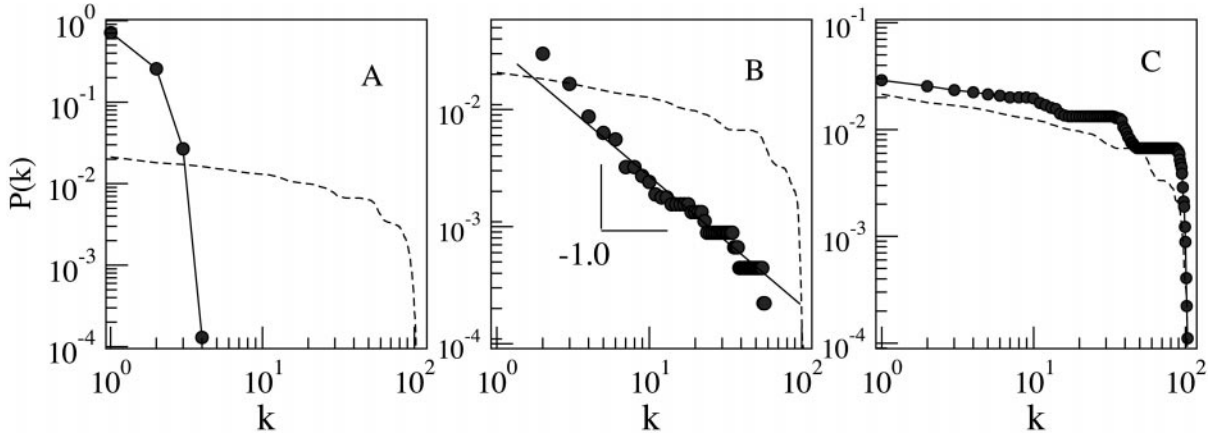

Fig. 3. Signal normalized frequency, $P(k)$, versus rank, k, for $\lambda = 0.3$ (*A*), $\lambda = \lambda^* = 0.41$ (*B*), and $\lambda = 0.5$ (*B* and *C*) (averages over 30 replicas: $n = m = 150$ and $T = 2nm$). The dotted lines show the distribution that would be obtained if signals and objects connected after a Poissonian distribution of degrees with the same number of connections of the minimum energy configurations. The distribution in *B* is consistent with human language ($\alpha = 1$).

maximal values of $I_n(\mathcal{S}, \mathcal{R})$, making use of fast memory access and the (theoretically) unlimited memory storage of computers (28, 29).

$\lambda > \lambda^*$ implies not taking into account the effort of the speaker. Getting the right word for a specific object may become unaffordable beyond a certain vocabulary size. Furthermore, a one-to-one map implies that the number of signals has to grow accordingly as the number of objects to describe increases (when $m \to \infty$) and leads to a referential catastrophe. A referential catastrophe is supported by the statistics of human–computer interactions, where the largest vocabularies follow Zipf's law (30) and are associated with a higher degree of expertise of the computer user. As the repertoire of potential signals is exhausted, strategies based on the combination of simple units are encouraged. Such a catastrophe could have motivated word formation from elementary syllables or phonemes but also syntax through word combinatorics. In a different context, some authors have shown that natural selection favors word formation or syntax when the number of required signals exceeds a threshold value (6). We show that arranging signals according to Zipf's law is the optimal solution for maximizing the referential power under effort for the speaker constraints. Moreover, almost the best $I_n(\mathcal{S}, \mathcal{R})$ is achieved before being forced to use one-to-one signal–object maps (Fig. 2). Although other researchers have shown how overcoming phase transitions could have been the origin of the emergence of syntax (5), our results suggest that early human communication could have benefited from remaining in a referential phase transition. There, communication is optimal with regard to the tradeoff between speaker and hearer needs. An evolutionary prospect is that the number of objects to describe can grow, keeping the size of the lexicon relatively small at the transition.

Having determined the only three optimal configurations resulting from tuning speaker and hearer requirements, the path toward human language can be traced hypothetically: (*i*) a transition from a no-communication phase ($\lambda < \lambda^*$) to a perfect-communication phase providing some kind of rudimentary referential signaling ($\lambda < \lambda^*$); (*ii*) a transition from a communication phase to the edge of the transition ($\lambda = \lambda^*$), where vocabularies can grow affordably (in terms of the speaker's effort) when $m \to \infty$. The latter step is motivated by the positive correlation between brain size and cognitive skills in primates (where m can be seen as a simple measure of them) (31). Humans may have had a pressure for economical signaling systems (given by large values of m) that other species did not have. The above-mentioned emergence of Zipf's law in the usage of computer commands (the only evidence known of evolution toward Zipf's law, although the context is not human–human interactions) is associated with larger repertoires (30), suggesting that there is a minimum vocabulary size and a minimum number of objects encouraging Zipf's law arrangements.

The relationship between both is straightforward if the hearer imposes its needs, because the number of signals must be exactly the number of objects (when $n = m$) in that case. Our results predict that no natural intermediate communication system can be found between small-sized lexica and rich lexica unless Zipf's law is used (Fig. 2*B*). This might explain why human language is unique with regard to other species but not only so. One-to-one maps between signals and objects are the distinguishing feature of index reference (2). Symbolic communication is a higher-level reference in which reference results basically from interactions between signals (2). Zipf's law appears on the edge of the indexical communication phase and implies polysemy. The latter is the necessary (but not sufficient) condition for symbolic reference (2). Our results strongly suggest that Zipf's law is required by symbolic systems.

We thank P. Fernández, R. Köhler, P. Niyogi, and M. Nowak for helpful comments. This work was supported by the Institució Catalana de Recerca i Estudis Avançats, the Grup de Recerca en Informàtica Biomèdica, the Santa Fe Institute (to R.V.S.), Generalitat de Catalunya Grant FI/2000-00393 (to R.F.i.C.), and Ministerio de Ciencia y Technologia Grant BFM 2001-2154 (to R.V.S.).

1. Chomsky, N. (1968) *Language and Mind* (Harcourt, Brace, and World, New York).
2. Deacon, T. W. (1997) *The Symbolic Species: The Co-evolution of Language and the Brain* (Norton & Company, New York).
3. Pinker, S. & Bloom, P. (1990) *Behav. Brain Sci.* **13,** 707–784.
4. Bickerton, D. (1990) *Language and Species* (Chicago Univ. Press, Chicago).
5. Nowak, M. A. & Krakauer, D. C. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 8028–8033.
6. Nowak, M. A., Plotkin, J. B. & Jansen, V. A. (2000) *Nature* **404,** 495–498.
7. Hauser, M. D. (1996) *The Evolution of Communication* (MIT Press, Cambridge, MA).
8. Miller, G. (1981) *Language and Speech* (Freeman, San Francisco).
9. Ujhelyi, M. (1996) *J. Theor. Biol.* **180,** 71–76.
10. Köhler, R. (1987) *Theor. Linguist.* **14,** 241–257.
11. Zipf, G. K. (1949) *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology* (Addison–Wesley, Cambridge, MA); reprinted in Zipf, G. K. (1972) *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology* (Hafner, New York), 1st ed., pp. 19–55.
12. Gernsbacher, M. A., ed. (1994) *Handbook of Psycholinguistics* (Academic, San Diego).
13. Köhler, R. (1986) *Zur Linguistischen Synergetik: Struktur und Dynamik der Lexik* (Brockmeyer, Bochum, Germany).
14. Balasubrahmanyan, V. K. & Naranan, S. (1996) *J. Quant. Linguist.* **3,** 177–228.
15. Ash, R. B. (1965) *Information Theory* (Wiley, New York).
16. Solé, R. V., Manrubia, S. C., Luque, B., Delgado, J. & Bascompte, J. (1996) *Complexity* **1,** 13–26.
17. Binney, J., Dowrick, N., Fisher, A. & Newman, M. (1992) *The Theory of Critical Phenomena: An Introduction to the Renormalization Group* (Oxford Univ. Press, New York).
18. Miller, G. A. & Chomsky, N. (1963) in *Handbook of Mathematical Psychology*, eds. Luce, R. D., Bush, R. & Galanter, E. (Wiley, New York), Vol. 2.
19. Mandelbrot, B. (1966) in *Readings in Mathematical Social Sciences*, eds. Lazarsfield, P. F & Henry, N. W. (MIT Press, Cambridge, MA), pp. 151–168.
20. Simon, H. A. (1955) *Biometrika* **42,** 425–440.
21. Pietronero, L., Tosatti, E., Tosatti, V. & Vespignani, A. (2001) *Physica A* **293,** 297–304.
22. Nicolis, J. S. (1991) *Chaos and Information Processing* (World Scientific, Singapore).
23. Naranan, S. & Balasubrahmanyan, V. (1998) *J. Quant. Linguist.* **5,** 35–61.
24. Miller, G. A. (1957) *Am. J. Psychol.* **70,** 311–314.
25. Li, W. (1992) *IEEE Trans. Inf. Theor.* **38,** 1842–1845.
26. Cohen, A., Mantegna, R. N. & Havlin, S. (1997) *Fractals* **5,** 95–104.
27. Ferrer i Cancho, R. & Solé, R. V. (2002) *Adv. Complex Syst.* **5,** 1–6.
28. Steels, L. (1996) in *Proceedings of the 5th Artificial Life Conference*, ed. Langton, C. (Addison–Wesley, Redwood, CA).
29. Nowak, M. A., Plotkin, J. B. & Krakauer, D. C. (1999) *J. Theor. Biol.* **200,** 147–162.
30. Ellis, S. R. & Hitchcock, R. J. (1986) *IEEE Trans. Syst. Man Cybern.* **16,** 423–427.
31. Reader, S. M. & Laland, K. N. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 4436–4441.