The small world of human language

2 authors:

Ramon Ferrer-i-Cancho
Polytechnic University of Catalonia
170 PUBLICATIONS   **7,605** CITATIONS

SEE PROFILE

Ricard Sole
Pompeu Fabra University
425 PUBLICATIONS   **23,474** CITATIONS

SEE PROFILE

PROCEEDINGS OF **THE ROYAL SOCIETY** | **B** | **BIOLOGICAL SCIENCES**

The small world of human language

Ramon Ferrer i Cancho and Richard V. Solé

References	**Article cited in:** http://rspb.royalsocietypublishing.org/content/268/1482/2261#related-urls
Email alerting service	Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click **here**

The small world of human language

Ramon Ferrer i Cancho[1][*] **and Ricard V. Solé**[1,2]

[1]*Complex Systems Research Group, Department of Physics, Universitat Politècnica de Catalunya, Campus Nord B4, 08034 Barcelona, Spain*
[2]*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA*

Words in human language interact in sentences in non-random ways, and allow humans to construct an astronomic variety of sentences from a limited number of discrete units. This construction process is extremely fast and robust. The co-occurrence of words in sentences reflects language organization in a subtle manner that can be described in terms of a graph of word interactions. Here, we show that such graphs display two important features recently found in a disparate number of complex systems. (i) The so called small-world effect. In particular, the average distance between two words, d (i.e. the average minimum number of links to be crossed from an arbitrary word to another), is shown to be $d \approx 2-3$, even though the human brain can store many thousands. (ii) A scale-free distribution of degrees. The known pronounced effects of disconnecting the most connected vertices in such networks can be identified in some language disorders. These observations indicate some unexpected features of language organization that might reflect the evolutionary and social history of lexicons and the origins of their flexibility and combinatorial nature.

Keywords: small world; scale-free networks; lexical networks; human language

1. INTRODUCTION

The emergence of human language is one of the major transitions in evolution (Smith & Szäthmáry 1997). Living humans have a symbolic mind capable of language that is not shared by any other species. Over two million years of hominid evolution, a coevolutionary exchange between languages and brains took place (Deacon 1997). This process involved the (possibly sudden) transition from non-syntactic to syntactic communication (Nowak & Krakauer 1999; Nowak *et al.* 2000). Human language allows the construction of a virtually infinite range of combinations from a limited set of basic units. The process of sentence generation is astonishingly rapid and robust, and indicates that we are able to rapidly gather words to form sentences in a highly reliable fashion.

A complete theory of language requires a theoretical understanding of its implicit statistical regularities. The best known of them is the Zipf's law, which states that the frequency of words decays as a power function of its rank (Zipf 1972). However, in spite of its relevance and universality (Balasubrahmanyan & Naranan 1996), such a law can be obtained by various mechanisms (Nicolis 1991; Simon 1955; Li 1992) and does not provide deep insight into the organization of language. The reason is that information transmission is organized into sentences that are made by words in interaction with each other.

Human brains store lexicons that are usually formed by thousands of words. Estimates are in the range 10^4-10^5 words (Romaine 1992; Miller & Gildea 1987). Besides, the contents of the lexicon of individuals using the same language vary depending on many factors, such as age, geographical location, social context, education and profession. The primary goal of a lexicon is to achieve successful communication, so a common lexicon must exist for successful basic communication between speakers, hereafter named a kernel lexicon, to surmount the limitations imposed by the factors mentioned above. Obviously, the best candidates to form this lexicon are the most frequently used words. Actually, the analysis of multi-speaker collections of texts (corpus) shows two different regimes that divide words into basic and specialized words (Cancho & Solé 2000).

Words interact in many ways. Some words co-occur with certain words at a higher probability than with others and co-occurrence is not trivial, i.e. it is not a straightforward implication of the known frequency distribution of words. If a text is scrambled, the frequency distribution is maintained but its content will not make sense.

In this paper, we show that the co-occurrence of words in sentences relies on the network structure of the lexicon, the properties of which are analysed in depth. As we will show in this paper, human language can be described in terms of a graph of word interactions. This graph has some unexpected properties (shared by other biological and technological networks (Amaral *et al.* 2000; Strogatz 2001)) that might underlie its diversity and flexibility, and create new questions about its origins and organization.

2. GRAPH PROPERTIES OF HUMAN LANGUAGE

Words co-occur in sentences. Many co-occurrences are due to syntactical relationships between words (e.g. head-modifier or dependency relationships (Melčuck 1989)). Some others are due to stereotyped expressions or collocations, the words of which work together (e.g. take it easy, New York). We will define links as significative co-occurrences between words in the same sentence. We do not seek to provide a detailed list of the origins and linguistic interpretation of such significative co-occurrences, but to show simply that they exist and can be captured using quantitative measures of correlation regardless of their nature. A first approach for estimating the network of the lexicon is to consider that there is a link between every pair of neighbouring words (at the risk of capturing spurious correlations).

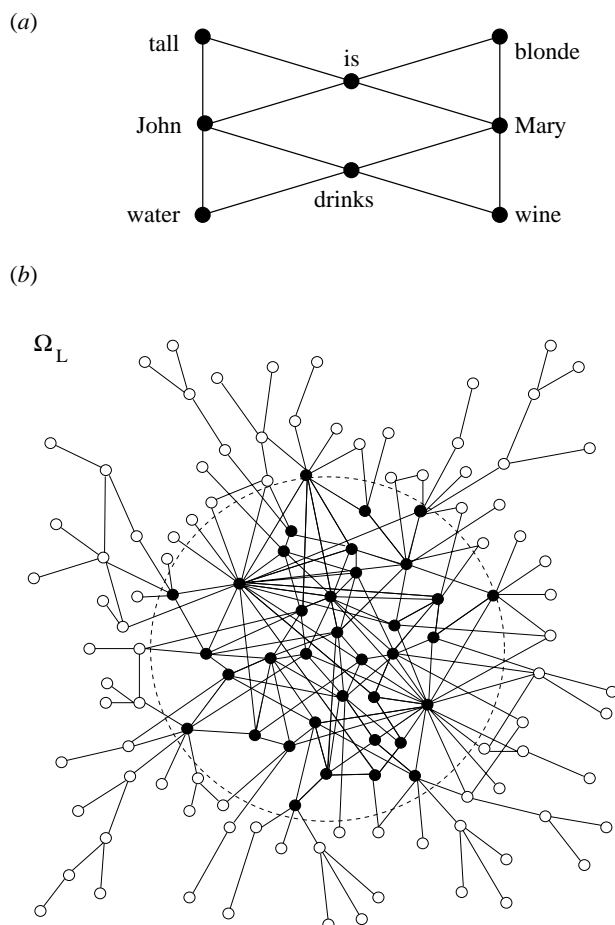*Author for correspondence (ramon@complex.upc.es).

(a)

(b)

Figure 1. Word networks. (a) A toy network constructed with four sentences: 'John is tall.' 'John drinks water.' 'Mary is blonde.' 'Mary drinks wine.' (b) A possible pattern of wiring in Ω_L. Black nodes are common words and white nodes are rare words. Two words are linked if they co-occur significantly.

The most correlated words in a sentence are the closest. A decision must be taken about the maximum distance considered for forming links. The top of figure 1 shows a simple (toy) graph that has been constructed linking words at a distance of one or two in the same sentence. If the distance is long, the risk of capturing spurious co-occurrences increases. If the distance is too short, certain strong co-occurrences can be systematically not taken into account. We decided the maximum distance according to the minimum distance at which most of the co-occurrences are likely to happen.

(i) Many co-occurrences take place at a distance of one, e.g. 'red flowers' (adjective–noun), 'the/this house' (article–determiner-noun), 'stay here' (verb–adverb), 'getting dark' (verb–adjective),'can see' (modal–verb).
(ii) Many co-occurrences take place at a distance of two, e.g. 'hit the ball' (verb–object), 'Mary usually cries' (subject–verb), 'table of wood' (noun–noun through a prepositional phrase), 'live in Boston' (verb–noun).

Long-distance correlations, i.e. at a distance greater than two, have been shown to take place in human sentences (Chomsky 1957). Here, we stop our seek at a distance of two. The reason is fourfold.

(i) Consideration of any distance requires an automatic procedure for accomplishing the task of capturing the relevant links. We do not know of any computational technique that successfully carries out this task for a general case. From a practical point of view, a context of two words is considered to be the lowest distance at which most of the improvement of certain computational linguistics methods is achieved (Kaplan 1955; Choueka & Lusignan 1985).
(ii) Our method fails to capture the exact relationships that happen in a particular sentence, but does capture (almost) every possible type of link. The type of link is determined by the syntactic categories or roles of the intervening words. Very few types of link (if any) are observed at a distance greater than two and none at lower distances.
(iii) We are not interested in all the relationships that happen in a particular sentence. Our goal is to capture as many links as possible through an automatic procedure. If the corpus is big enough, the macroscopic properties of the network should emerge.
(iv) Being syntactic dependencies non-crossing (Hudson 1984; Melčuck 1989), a long-distance syntactic link implies the existence of lower-distance syntactic links. By contrast, a short-distance link does not imply a long-distance link.

The technique can be improved by choosing only pairs of consecutive words, the mutual co-occurrence of which is larger than expected by chance. This can be measured with the condition $p_{ij} > p_i p_j$, which defines the presence of correlations beyond that expected from a random ordering of words. If a pair of words co-occurs less than expected when independence between such words is assumed, the pair is considered to be spurious. Graphs in which this condition is used will be called 'restricted' ('unrestricted' otherwise). Punctuation marks are skipped during the processing of the corpus.

Let us consider the graph of human language, Ω_L, as defined by $\Omega_L = (W_L, E_L)$, where $W_L = \{w_i\}$, $(i = 1, \ldots, \mathcal{N}_L)$ is the set of $\mathcal{N}_L$ words and $E_L = \{\{w_i, w_j\}\}$ is the set of edges or connections between words. Here, $\xi_{ij} = \{w_i, w_j\}$ indicates that there is an edge (and thus a link) between words w_i and w_j. Two connected words are adjacent and the degree of a given word is the number of edges that connects it with other words. The bottom of figure 1 shows what such a network would look like.

Recent research on a number of biological, social and technological graphs showed that they share a common feature: the so-called small-world (SW) property (Watts & Strogatz 1998). SW graphs have a number of surprising properties that make them specially relevant to understanding how interactions among individuals, metabolites or species lead to the robustness and homeostasis observed in nature (Watts & Strogatz 1998). The SW pattern can be detected from the analysis of two basic statistical properties: the so-called clustering coefficient C and the path length d. Let us consider the set of links ξ_{ij} $(i, j = 1, \ldots, \mathcal{N}_L)$, where $\xi_{ij} = 1$ if a link exists and 0 otherwise and that the average number of links per word is $\bar{k}$. Let us indicate by $\Gamma_i = \{j | \xi_{ij} = 1\}$ the set of nearest neighbours of a word $w_i \in W_L$. The clustering coefficient

for this word is defined as the number of connections between the words $w_j \in \Gamma_i$. By defining

$$\mathcal{L}_i = \sum_{j=1}^{\mathcal{N}_L} \xi_{ij} \left[\sum_{k \in \Gamma_i; j < k} \xi_{jk} \right], \tag{2.1}$$

we have

$$C_v(i) = \frac{\mathcal{L}_i}{\binom{|\Gamma_i|}{2}}.$$

Therefore, the clustering coefficient is the average over W_L:

$$C = \frac{1}{\mathcal{N}_L} \sum_{i=1}^{\mathcal{N}_L} C_v(i) \tag{2.2}$$

and measures the average fraction of pairs of neighbours of a node that are also neighbours of each other.

The second measure is easily defined. Given two words $w_i, w_j \in W_L$, let $d_{\min}(i, j)$ be the minimum path length that connects these two words in Ω_L. The average path length of a word will be defined as

$$d_v(i) = \frac{1}{\mathcal{N}_L} \sum_{j=1}^{\mathcal{N}_L} d_{\min}(i, j), \tag{2.3}$$

and thus the average path length d will be

$$d = \frac{1}{\mathcal{N}_L} \sum_{i=1}^{\mathcal{N}_L} d_v(i). \tag{2.4}$$

Graphs with SW structure are highly clustered, but d is small. Random graphs (in which nodes are randomly wired) are not clustered and also have a short d (Watts & Strogatz 1998). At the other extreme, regular lattices with only nearest-neighbour connections among units are typically clustered and show long paths. It has been shown, however, that a regular lattice can be transformed into a SW if a small fraction of nodes are rewired to randomly chosen nodes. Thus, a small degree of disorder generates short paths (as in the random case) but preserves the local neighbourhood (Watts & Strogatz 1998).

For random graphs, $C_v^{rand} \approx \bar{k}/\mathcal{N}$. For SW graphs, d is close to that expected for random graphs, d^{rand}, with the same $\bar{k}$ and $C_v \gg C_v^{rand}$. These two conditions are taken as the standard definition of SW. SW graphs have been shown to be present in both social and biological networks (Jeong *et al.* 2000; Montoya & Solé 2001; Strogatz 2001; Amaral *et al.* 2000). Besides, some of these networks also show scaling in their degree distribution. In other words, the probability $P(k)$ of having a node with degree k scales as $P(k) \approx k^{-\gamma}$. We have found that the graph of human language displays similar properties. This second property has been shown to be related to an extremely high stability against perturbations directed to randomly chosen nodes and a high fragility when perturbations are directed to highly connected ones (Albert *et al.* 2000). As we will show here, Ω_L shows both SW structure and a power law in $P(k)$.

Table 1. Word network patterns.

(It can be seen that $C \gg C_{\text{random}}$ and $d \approx d_{\text{random}}$, consistently in a SW network. All values are exact except for those marked with an asterisk, which have been estimated on a random subset of the vertices (after having processed 2% of the vertices, fluctuations in d^* as a function of the subset size clearly affected only the third decimal digit).)

graph	C	C_{random}	d	d_{random}
Ω_L (UWN)	0.687	1.55×10^{-4}	2.63^*	3.03
Ω_L (RWN)	0.437	1.55×10^{-4}	2.67^*	3.06

3. SCALING AND SMALL-WORLD PATTERNS

The biggest connected component of the networks that results from the basic and improved methods will be called, respectively, the unrestricted word network (UWN) and the restricted word network (RWN). They have $\mathcal{N}(\text{UWN}) = 478\,773$ and $\mathcal{N}(\text{RWN}) = 460\,902$ nodes, with $E(\text{UWN}) = 1.77 \times 10^7$ and $E(\text{RWN}) = 1.61 \times 10^7$ edges, respectively. With average connectivities of $\bar{k}_{\text{uwn}} = 74, 2$ and $\bar{k}_{\text{rwn}} = 70, 13$, their clustering and path lengths are indicated in Table 1.

Figure 2 shows the distribution of degrees of both the UWN and RWN obtained after processing about three-quarters of the 10^7 words of the British National Corpus (http://info.ox.ac.uk/bnc/). The obvious limitations of our methods are overcome by the use of a large amount of data. The distribution of connectivities of UWN and RWN decays with two different average exponents each, $\gamma_1 = -1.50$ for the first regime and $\gamma_2 = -2.70$ for the second regime. The exponent in the second regime is similar to that of the so-called Barabási-Albert (BA) model ($\gamma_{\text{BA}} = -3$) (Barabási & Albert 1999). The BA model leads to scale-free distributions using the rule of preferential attachment. The rule simply assumes that new nodes in the growing network are preferentially attached to an existing node with a probability proportional to the degree of such a node.

Furthermore, word networks have SW features. The average minimum distance between vertices is below 3 (2.63 for the UWN and 2.67 for the RWN), so reaching whatever vertex involves fewer than three jumps on average. This is significant, because the network contains about 4.7×10^5 different words. Clustering (0.687 for the UWN and 0.437 for the RWN) is far from the random expectation (1.55×10^{-4} for both the UWN and the RWN) in both cases.

As far as we know, this is the first time that such a statistically significant property has been reported about the organization of human language. In spite of the huge number of words that can be stored by a human, any word in the lexicon can be reached with fewer than three intermediate words, on average. If a word is reached during communication, jumping to another word requires very few steps. Speed during speech production is important and can be more easily achieved if intervening words are close to each other in the underlying structure used for the construction of sentences. Conversely, richness is another quality of a powerful communication. Although words are preferably chosen from the kernel lexicon,
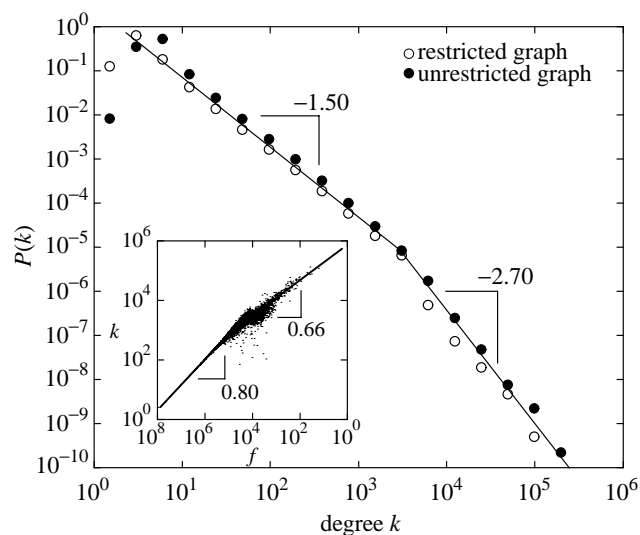
Figure 2. Degree distribution for the unrestricted word network (filled circles) and the restricted word network (open circles). Points are grouped by powers of two. Inset: average degree as a function of frequency. Degree increases as a function with frequency, with exponent 0.80 for the first domain and 0.66 for the second one.
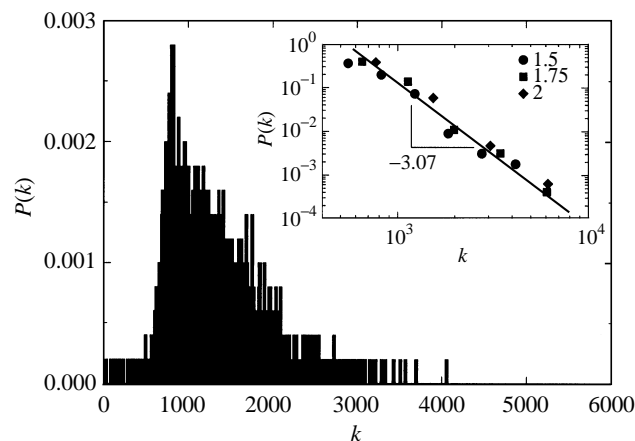

Figure 3. Connectivity distribution for the kernel word network, formed by the 5000 most connected vertices in RWN. Inset: power-law tail for $k > \bar{k}$ calculated by grouping in powers of 1.5, 1.75 and 2. The exponent of the power tail is $\gamma_{KWN} \approx -3$, indicating that preferential attachment is happening.

external words are at a short distance, so rich communication based on the word network can be attained with little increase in effort.

It is well known that the more frequent a word, the more available it is for production (Brown & McNeil 1966; Kempen & Hijbers 1983) and comprehension (Forster & Chambers 1973; Scarborough *et al.* 1977) processes. This phenomenon is known as the frequency (referring to the whole individual's experience) or recency (referring to the recent individual's experience) effect (Akmajian 1995). This phenomenon shows that preferential attachment is very likely to shape the scale-free distribution of degrees in a similar way to the BA model. For the most frequent words, $k \propto f^{0.66}$, where k is the degree and f is the frequency of the word. We can then recast the frequency effect in terms of the degree as 'the higher the degree of a word, the higher its availability'. In other words, links that include highly connected words are preferred. The inset in figure 2 shows the complete relationship between f and k in RWN.

The exponent of UWN and RWN is closer to $\gamma_{BA} = -3$ in the second regime of the distribution, which is where the frequency effect makes much more sense. The kernel lexicon contains words that are common to the whole community of speakers and its size is determined as the rank at which there is a change in the exponents of the word frequency versus rank plot (Cancho & Solé 2000). Beyond the kernel, a certain word is unknown for one speaker and familiar for another. The frequency and recency effects then cannot be applied for all the individuals that contribute to shaping the underlying lexicon network. It is thus expected that the network formed exclusively by the interaction of kernel words, hereafter named the kernel word network (KWN), better agrees with the predictions that can be performed when preferential attachment is at play. Figure 3 shows the log-normal appearance of the connectivity distribution. The power tail has exponent $\gamma_{KWN} \approx -3$, consistent with the BA

model and the differences with respect to it require special attention. It is important to note that the kernel lexicon is a versatile subset of the repertoire of individual speakers. A few thousand words must be able to say everything or almost everything. Even when lexicons become very small, i.e. pidgin languages, the lexicons of which do not usually exceed about 1000 words (Romaine 1992), it has been pointed out that they can say everything that can be said in a complex lexicon (e.g. English) at the expense of high redundancy (recurring to circumlocution). The average connectivity in the kernel is $\bar{k} = 1219$. A first consequence is that words with low connectivity must be rare. Having rather useless words in this crucial subset is an enormous waste. Once connected words become frequent in the distribution, the network organizes in a scale-free way. We believe that the scale-freeness is responsible for the ability to say everything of the kernel. A non-trivial network is needed because every word on average is connected to 24% of the rest of the kernel words.

4. DISCUSSION

We have shown that the graph connecting words in language shows the same statistical features as other complex networks. The short distance between words arising from the SW structure indicates that language evolution might have involved the selection of a particular arrangement of connections between words. Future work should theoretically address this problem, perhaps using an evolutionary language game model (Nowak & Krakauer 1999; Nowak *et al.* 2000) in which a pay-off associated to the graph structure is introduced. Concerning the scaling in $P(k)$ and the observed exponents, this pattern also calls for an evolutionary explanation. The word network is the result of a growth process in which new words are added and are likely to be linked to highly connected existing words.

If the SW features derive from optimal navigation needs, two predictions can be formulated. First, words the main purpose of which is to speed-up navigation must exist. Second, deriving from the first, brain disorders

characterized by navigation deficits in which such words are involved must exist. The best candidates for answering the first question are the so-called particles, a subset of the function words (e.g. articles, prepositions and conjunctions) formed by the most frequent among them (e.g. and, the, of). According to our calculations, the 10 most connected words are 'and', 'the', 'of', 'in', 'a', 'to', ''s', 'with', 'by' and 'is'. These words are characterized by a very low or zero semantic content. Although they are supposed to contribute to the sentence structure, they are not generally crucial for sentence understanding. A compelling test of this statement is that particles are the first words to be suppressed in telegraphic speech (Akmajian 1995).

The answer to the second prediction is agrammatism, a kind of aphasia in which speech is non-fluent, laboured, halting and lacking in function words (and thus of particles). Agrammatism is the only syndrome in which function words are particularly omitted (Caplan 1994). Function words are the most connected ones. We suggest that such halts and lack of fluency are due to fragility associated with the removal of highly connected words. Although scale-free networks are very tolerant to random removal of vertices, if deletion is directed to the most connected vertices the network gets broken into pieces (Albert *et al*. 2000).

It is known that the omission of function words is often accompanied by substitution of such words. Patients in which substitutions predominate and speech is fluent are said to undergo paragrammatism (Caplan 1994). We suggest that paragrammatism recovers fluency (i.e. low average word–word distance) by inadequately using the remaining highly connected vertices and thus often producing substitutions of words during discourse.

REFERENCES

Akmajian, A. 1995 *Linguistics. An introduction to language and communication*, ch. 2. Cambridge, MA: MIT Press.

Albert, R., Jeong, H. & Barabási, A.-L. 2000 Error and attack tolerance of complex networks. *Nature*, **406**, 378–381.

Amaral, L. A. N., Scala, A., Barthélémy, M. & Stanley, H. E. 2000 Classes of behaviour of small-world networks. *Proc. Natl Acad. Sci. USA* **97**, 11 149–11 152.

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

Barabási, A.-L. & Albert, R. 1999 Emergence of scaling in random networks. *Science* **286**, 509–511.

Brown, R. & McNeil, D. 1966 The 'tip of the tongue' phenomenon. *J. Verbal Learn. Verbal Behav.*, **5**, 325–337.

Cancho, R. F. i & Solé, R. V. 2000 Two regimes in the frequency of words and the origin of complex lexicons: Zipf's law revisited. Santa Fe Working Paper 00-12-068. *J. Quantitative Linguistics*. (Submitted.)

Caplan, D. 1994 *Language. Structure, processing and disorders*. MIT Press.

Chomsky, N. 1957 *Syntactic structures*. The Hague: Mouton.

Choueka, Y. & Lusignan, S. 1985 Disambiguation by short contexts. *Comp. Humanities* **19**, 147–157.

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

Forster, K. I. & Chambers, S. M. 1973 Lexical access and naming time. *J. Verbal Learn. Verbal Behav.* **12**, 627–635.

Hudson, R. 1984 *Word grammar*. Oxford, UK: B. Blackwell.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A.-L. 2000 The large-scale organization of metabolic networks. *Nature* **407**, 651–654.

Kaplan, A. 1955 An experimental study of ambiguity and context. *Mechanical Translation* **2**, 39–46.

Kempen, G. & Hijbers, P. 1983 The lexicalization process in sentence production and naming: indirect election of words. *Cognition* **14**, 185–209.

Li, W. 1992 Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Trans. Inform. Theory* **38**, 1842–1845.

Melčuck, I. 1989 *Dependencies grammar: theory and practice*. New York: University of New York.

Miller, G. A. & Gildea, P. M. 1987 How children learn words. *Scient. Am.* **257**, 94–99.

Montoya, J. M. & Solé, R. V. 2001 Small world patterns in food webs. Santa Fe Working Paper 00-10-059. *J. Theor. Biol.* (In the press.)

Nicolis, J. S. 1991 *Chaos and information processing*. Singapore: World Scientific.

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

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

Romaine, S. 1992 The evolution of linguistic complexity in pidgin and creole languages. In *The evolution of human languages* (ed. J. A. Hawkins & M. Gell-Mann), pp. 213–238. Redwood City, CA: Addison Wesley.

Scarborough, D. L., Cortese, C. & Scarborough, H. S. 1977 Frequency and repetition effects in lexical memory. *J. Exp. Psychol. Hum. Percept. Perform.* **3**, 1–7.

Simon, H. A. 1955 On a class of skew distribution functions. *Biometrika* **42**, 425–440.

Smith, J. M. & Szäthmáry, E. 1997 *The major transitions in evolution*. Oxford University Press.

Strogatz, S. H. 2001 Exploring complex networks. *Nature* **410**, 268–276.

Watts, D. J. & Strogatz, S. H. 1998 Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442.

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