THE ROYAL SOCIETY

doi 10.1098/rspb.2001.1824

Language as an evolving word web

S. N. Dorogovtsev[1,2*] and J. F. F. Mendes[1]

[1]*Departamento de Física and Centro de Física do Porto, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre 687, 4169-007 Porto, Portugal* (*sdorogov@fc.up.pt*)

[2]*AF Ioffe Physico-Technical Institute, 194021 St Petersburg, Russia* (*jfmendes@fc.up.pt*)

Human language may be described as a complex network of linked words. In such a treatment, each distinct word in language is a vertex of this web, and interacting words in sentences are connected by edges. The empirical distribution of the number of connections of words in this network is of a peculiar form that includes two pronounced power-law regions. Here we propose a theory of the evolution of language, which treats language as a self-organizing network of interacting words. In the framework of this concept, we completely describe the observed word web structure without any fitting. We show that the two regimes in the distribution naturally emerge from the evolutionary dynamics of the word web. It follows from our theory that the size of the core part of language, the 'kernel lexicon', does not vary as language evolves.

Keywords: evolution of language; word web; interaction of words; kernel lexicon

1. INTRODUCTION

How language evolves is a major challenge for linguistics and evolutionary biology (Smith & Száthmáry 1997; Deacon 1997; Hurford *et al.* 1998) and an intriguing problem for other sciences (Simon 1955, 1957; Nowak & Krakauer 1999; Nowak 2000; Nowak *et al.* 2000, 2001). The recent explosion of interest in networks (Albert *et al.* 1999; Barabási & Albert 1999; Huberman & Adamic 1999; Watts 1999; Strogatz 2001), including the World Wide Web and Internet (Huberman *et al.* 1998; Lawrence & Giles 1998, 1999; Albert & Barabási 2000), biological networks (Jeong *et al.* 2000), social (Watts & Strogatz 1998) and ecological webs (Williams & Martinez 2000), networks of collaborations (Newman 2001), etc., had an immediate consequence—the treatment of human language as a complex network of distinct words (Ferrer & Solé 2001, 2002).

This word web is arranged in the following way. The vertices of the web are the distinct words of language, and the undirected edges are connections between interacting words. It is not so easy to define the notion of word interaction in a unique way. Nevertheless, different reasonable definitions provide very similar structures of the word web. For instance, one can connect the nearest neighbours in sentences. Without going into detail, this means that the edge between two distinct words of language exists if these words are the nearest neighbours in at least one sentence in the bank of language. In such a definition, a pair of words may be connected via only one link, and multiple links are absent. One also may connect the second nearest neighbours and account for other types of the correlations between words. In fact, such a linking indicates the co-occurrence of words in sentences. It should be pointed out that the number of the connections of a word in the word web does not relate directly to the frequency of the occurrence of this word in language.

Recently it was found that this network has a complex architecture (Ferrer & Solé 2001, 2002) that differs dramatically from classical random graphs extensively studied in mathematical graph theory. In the papers of Ferrer & Solé (2001, 2002) the basic informative characteristic of the word web, i.e. the distribution of the numbers of connections of words, has been obtained empirically. In graph theory, the number of connections of a vertex is called its degree. The observed degree distribution of the word web has a long tail—unlike the Poisson degree distribution for the classical random graphs. This indicates that the word web belongs to the same class as the World Wide Web and the Internet (Albert *et al.* 1999; Huberman & Adamic 1999).

Moreover, the empirical degree distribution obtained by Ferrer & Solé (2001, 2002) has a complex form. It consists of two power-law parts with different exponents. This hampers any treatment but, however, makes it possible to find an explanation of the basic structure of the word web in the framework of a general concept. Indeed, if one proposes a theory which, without fitting, describes the empirical degree distribution and reproduces the values of all the characteristic scales, the announced aim will be achieved (it is hardly possible to describe such a complex form perfectly by coincidence). Here we present the solution of this problem.

2. THE MODEL

Human language is certainly an evolving system. Its present structure is determined by its past evolution. This system is so complex that it cannot be controlled but rather organizes itself while growing. We treat language as a growing network of interacting words. At its birth, a new word already interacts (collaborates) with several old ones. New interactions between old words emerge from time to time, and new edges arise.

How do words find their collaborators in language? Here we use the idea of preferential linking (preferential attachment of new edges to vertices with higher numbers of connections) (Barabási & Albert 1999). This fruitful idea is a particular realization of the general concept of Simon (1955, 1957). The simplest linear form of the preferential linking provides the power-law degree distributions for nets in which the average number of connections per

*Author for correspondence.

Received 30 May 2001 *Accepted* 7 August 2001

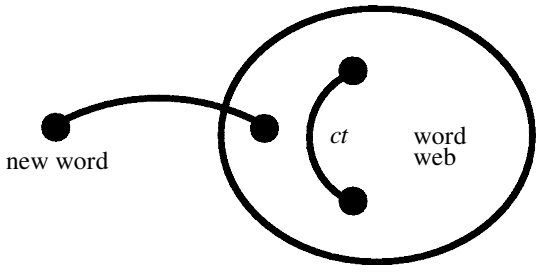

Figure 1. Scheme of the word web growth. At each time-step a new word appears, so t is the total number of words. The new word connects to some preferentially chosen old word. Simultaneously, ct new edges emerge between pairs of preferentially chosen old words. All the edges are undirected. We use the simplest kind of preferential attachment when a node is chosen with the probability proportional to the number of its connections.

vertex (the average degree) does not change during growth (Barabási & Albert 1999; Dorogovtsev *et al.* 2000; Krapivsky *et al.* 2000). If the total number of connections increases more rapidly than the number of vertices and the average degree grows, the exponent of the degree distribution takes a different value (Dorogovtsev & Mendes 2001). For the explanation of the resulting structure of the word web, we combine these two processes of edge emergence.

We use the following rules of the network growth (see figure 1). At each time-step, a new vertex (word) is added to the network, and the total number of vertices, t, plays the role of 'time'. We emphasize that t is not a real time but the size of the word web. At its birth, the new word connects to several old ones. We do not know the original number of connections. We only know that it is of the order of unity. It would be unfair to play with an unknown parameter to fit the experimental data, so we set this number to unity. (One may check that the introduction of this parameter does not noticeably change the degree distribution of the word web.) We use the simplest natural version of the preferential linking, so a new word is connected to some old one i with the probability proportional to its degree k_i, as in the Barabási–Albert model (Barabási & Albert 1999). In addition, at each increment of time, ct new edges emerge between old words, where c is a constant coefficient that characterizes a particular network. The linear dependence appears if each vertex makes new connections at a constant rate, so that we choose it as the most simple and natural. Notice that a similar linear dependence was observed in real collaboration networks (Barabási *et al.* 2001). These new edges emerge between old words i and j with the probability proportional to the product of their degrees $k_i k_j$ (Albert & Barabási 2000; Dorogovtsev & Mendes 2000). A very similar model, based on preferential attachment, was recently applied to the description of networks of collaborations of the co-authors of scientific publications (Barabási *et al.* 2001), so that our concept indicates the intimate relationship between the word web and collaboration networks.

Two slightly different methods (two distinct definitions of the pairwise word–word interactions) were used by Ferrer & Solé (2001, 2002) to construct the word web. The two resulting webs, obtained after processing three-quarters of a million words of the British National Corpus (a collection of text samples of both spoken and written modern British English), have nearly the same degree distributions, and each one contains about 470 000 vertices. The average number of connections per word (the average degree) is $\overline{k} \approx 72$. These are the only parameters of the word web that we know and can use in the model.

This stochastic model can be solved exactly, but here, for a simple presentation, we use the continuum approximation. Such an approach was proved to describe quite well the degree distributions of networks growing under the mechanism of preferential linking (Barabási & Albert 1999; Dorogovtsev & Mendes 2000, 2001). In our case, it provides the non-stationary degree distribution $P(k, t)$ very close to the exact one everywhere except for the narrow region $k<10$. One should emphasize that the continuum approach yields the exact values of the exponents of the distribution.

In the continuum approximation, the degrees of the vertices born at time s and observed at time t are substituted by their average value $k(s, t)$. For the large network, the evolution of $k(s, t)$ is described by the simple equation

$$\frac{\partial k(s,t)}{\partial t} = (1 + 2ct)\frac{k(s,t)}{\int_0^t \mathrm{d}u\, k(u,t)}, \tag{2.1}$$

with the obvious boundary condition $k(t, t) = 1$. The nature of this equation can be easily understood. The ratio on the right hand side is a direct consequence of the preferential attachment. At each time-step, $1 + 2ct$ ends of new edges are distributed preferentially. Indeed, one such an end belongs to the edge coming from a new word, and the others are the ends of the ct new edges emerging between old words. Here, we have presented heuristic arguments, but equation (2.1) can be derived more strictly (Dorogovtsev & Mendes 2000).

One sees that the total degree of the word web is $\int_0^t \mathrm{d}u\, k(u, t) = 2t + ct^2$, so the average degree of the network is equal to $\overline{k}(t) = 2 + ct$. The present value of the average degree of the word web is close to 72; hence $1 \ll ct \approx 70$. The solution of equation (2.1) is of a singular form

$$k(s,t) = \left(\frac{ct}{cs}\right)^{1/2}\left(\frac{2+ct}{2+cs}\right)^{3/2}, \tag{2.2}$$

which indicates the presence of two distinct regimes in this problem. From equation (2.2), using the standard expression for the degree distribution, $P(k, t) = -[t\partial k(s, t)/\partial s]^{-1}|_{s=s(k,t)}$, we immediately obtain the non-stationary degree distribution

$$P(k,t) = \frac{1}{ct}\frac{cs(2+cs)}{1+2cs}\frac{1}{k}, \tag{2.3}$$

where $s = s(k, t)$ is the solution of equation (2.2).

One sees from equations (2.2) and (2.3) that this non-stationary distribution has two regions with different behaviours separated by the crossover point $k_{\mathrm{cross}} \approx \sqrt{ct}(2 + ct)^{3/2}$. The crossover moves in the direction of large degrees as the network grows. Below this point, the degree distribution is stationary, $P(k) \cong \frac{1}{2}k^{-3/2}$ (we use the fact that in the word web $ct \gg 1$). Above the

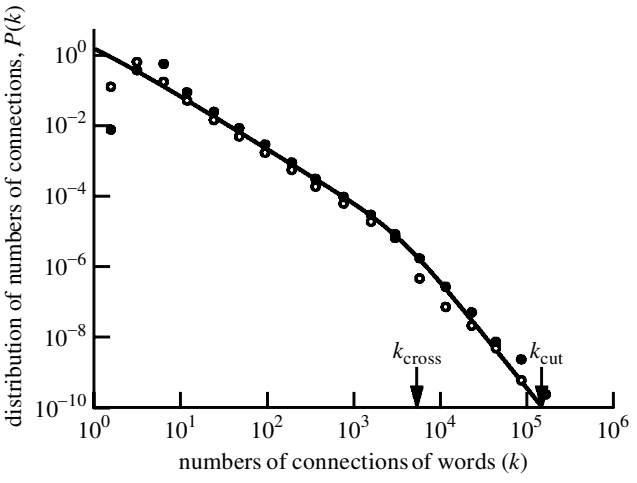

Figure 2. Distribution of the numbers of connections (degrees) of words in the word web on a log to log scale. The solid line is the result of our calculation using the parameters of the word web, the size $t \approx 470\,000$ and the average number of connections $\overline{k}(t) \approx 72$. Empty and filled circles show the distributions of the numbers of connections obtained by Ferrer & Solé (2001, 2002) for the two different methods of construction of the word web. In the region $k<10$, where the deviations of the continuous approximation from the exact solution of the model are noticeable, we present the exact solution. The arrows indicate the theoretically obtained point of crossover, k_{cross}, between the regions with the exponents 3/2 and 3 and the cut-off k_{cut} of the power-law dependence due to the size effect. For a better comparison, the theoretical curve is displaced upwards (note that the comparison is impossible in the region of the smallest k where the experimentally obtained distribution essentially depends on the definition of the word web).

crossover point, we obtain the behaviour $P(k,t) \cong \frac{1}{4}(t)^3 k^{-3}$, so that the degree distribution is non-stationary in this region. Thus, the model provides two distinct values for the degree distribution exponent, 3/2 for $k< k_{\mathrm{cross}}$ and 3 for $k>k_{\mathrm{cross}}$.

The degree distribution has one more important characteristic point, the cut-off produced by the size effect. Its position k_{cut} is easily estimated from the condition that only one vertex in the network is of degree exceeding k_{cut}, that is, $t\int_{k_{\mathrm{cut}}}^{\infty} \mathrm{d}k\, P(k) \sim 1$, and thus $k_{\mathrm{cut}} \sim \sqrt{t/8}(ct)^{3/2}$. Here we do not present the complete exact result that may be obtained using the master equation approach (Dorogovtsev *et al.* 2000). The infinite limit of the exact degree distribution takes the simple form $P(k, t \to \infty) = \frac{1}{2}B(k, 3/2)$ where $B(\ ,\ )$ is the β-function. Minor deviations from the continuum approximation are visible only for $k<10$.

In figure 2, we plot the degree distribution of the model (the solid line). To obtain the theoretical curve, we used equations (2.2) and (2.3) with the two known parameters of the word web. The deviations from the continuum approximation are accounted for in the small k region, $k<10$. One sees that the agreement with the empirical data (Ferrer & Solé 2001, 2002) is excellent. Note that we do not use any fitting. For a better comparison, in figure 2, the theoretical curve is displaced upward (we have to exclude two experimental points with the smallest k because these points are dependent on the method of the construction of the word web, and any comparison in this region is meaningless in principle).

From the relations obtained above, we find the characteristic values for the crossover and cut-off, $k_{\mathrm{cross}} \approx 5.1 \times 10^3$, that is, $\log_{10} k_{\mathrm{cross}} \approx 3.7$, and $\log_{10} k_{\mathrm{cut}} \approx 5.2$. From figure 2, one sees that these values coincide with the experimental ones. As far as we know, this is the first time that such complex empirical data for networks have been described without fitting. We should emphasize that the extent of agreement is truly surprising.

3. DISCUSSION

The minimal model does not account for numerous, at first sight, important factors, e.g. the death of words or the variations of words during the evolution of language. Note that the clustering coefficient of the word web takes large values (Ferrer & Solé 2001, 2002). One can easily explain this by incorporating the following features: (i) a word simultaneously makes not one but several connections; and (ii) these edges usually connect to already interacting words. Here we do not account for these circumstances because we are interested only in the degree distribution. One may check that a deviation of the number of connections of new-born words from unity does not change the exponents and inessentially shifts the crossover point.

The agreement is convincing because it is approached over the whole range of values of k, that is, over five decades. In fact, the word web turns out to be very convenient in this respect because the total number of edges in it is extremely high (about 3.4×10^7) and the value of the cut-off degree is large.

Notice that few words are in the region above the crossover point $k_{\mathrm{cross}} \approx 5.1 \times 10^3$. With the growth of language, k_{cross} increases rapidly but, as follows from our relations, the total number of words of degree greater than k_{cross} does not change. It is a constant of the order of $1/(8c) \approx t/(8\overline{k}) \sim 10^3$, i.e. of the order of the size of a small set of words forming the kernel lexicon of British English that was estimated as 5000 words (Ferrer & Solé 2001, 2002) and is the most important core part of language. Therefore, our concept suggests that the number of words in this part of language does not depend essentially on the total number of distinct words in language. Formally speaking, the size of the kernel lexicon is determined by the value of the average rate c with which words find new partners in language. The word web has been constructed only for British English, and the comparison of word webs of various languages is a challenge for the future.

There exist many obvious ways to improve the minimal model used above. Nevertheless, at present, such attempts seem rather meaningless because, as we have noted, it is hard to define rigorously the procedure of the word web construction, and the experimental data do not allow us to make a better comparison.

4. CONCLUSIONS

We have proposed a simple stochastic theory of evolution of human language based on the treatment of language as an evolving network of interacting (collaborating) words.

The structure of language is the result of the self-organization of the word web during its growth. The key result is the distribution of the number of connections of words. We have found that the self-organization produces the most connected small kernel lexicon of language, the size of which does not change essentially along the language evolution. The degree distribution of words in this core of language crucially differs from the degree distribution for the rest of language. We have shown that the basic characteristic of the word web structure, namely the degree distribution, does not depend on the rules of language but is determined by the general principles of the evolutionary dynamics of the word web. We would like to note that the successful description is important because recent progress in the understanding of numerous stochastic multiplicative processes in nature is based on the Simon model (Simon 1955, 1957) which was originally applied to the description of the structure of human language.

This work was supported by the project POCTI/1999/FIS/33141. We thank A. N. Samukhin for helpful discussions and G. Tripathy for reading our manuscript thoroughly.

REFERENCES

Albert, R. & Barabási, A.-L. 2000 Topology of evolving networks: local events and universality. *Phys. Rev. Lett.* **85**, 5234–5237.

Albert, R., Jeong, H. & Barabási, A.-L. 1999 Diameter of the World Wide Web. *Nature* **401**, 130–131.

Albert, R., Jeong, H. & Barabási, A.-L. 2000 Error and attack tolerance of complex networks. *Nature* **406**, 378–382.

Barabási, A.-L. & Albert, R. 1999 Emergence of scaling in random networks. *Science* **286**, 509–512.

Barabási, A.-L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A. & Vicsek, T. 2001 Evolution of the social network of scientific collaborations. See http://arxiv.org/abs/condmat/0104162.

Deacon, T. W. 1997 *The symbolic species: the coevolution of language and the brain*. New York: W. W. Norton.

Dorogovtsev, S. N. & Mendes, J. F. F. 2000 Scaling behaviour of developing and decaying networks. *Europhys. Lett.* **52**, 33–39.

Dorogovtsev, S. N. & Mendes, J. F. F. 2001 Effect of the accelerating growth of communications networks on their structure. *Phys. Rev.* E **63**, 025101 (R).

Dorogovtsev, S. N., Mendes, J. F. F. & Samukhin, A. N. 2000 Structure of growing networks with preferential linking. *Phys. Rev. Lett.* **85**, 4637–4640.

Ferrer, R. & Solé, R. V. 2000*a* Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *J. Quantitative Linguistics*. (In the press.)

Ferrer, R. & Solé, R. V. 2000*b* Santa Fe working papers 00-12-068. See http://www.santafe.edu/sfi/publications/abstracts/00-12-068abs.html.

Ferrer, R. & Solé, R. V. 2001 Santa Fe working papers 01-03-016. See http://www.santafe.edu/sfi/publications/abstracts/01-03-016abs.html.

Ferrer, R. & Solé, R. V. 2002 The small-world of human language. *Proc. R. Soc. Lond.* B. (Submitted.)

Huberman, B. A. & Adamic, L. A. 1999 Growth dynamics of the World-Wide Web. *Nature* **401**, 131.

Huberman, B. A., Pirollo, P. L. T., Pitkow, J. E. & Lukose, R. M. 1998 Strong regularities in world wide web surfing. *Science* **280**, 95–97.

Hurford, J. R., Studdert-Kennedy, M. & Knight, C. (eds) 1998 *Approaches to the evolution of language*. Cambridge University Press.

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

Krapivsky, P. L., Redner, S. & Leyvraz, F. 2000 Connectivity of growing random networks. *Phys. Rev. Lett.* **85**, 4633–4636.

Lawrence, S. & Giles, C. L. 1998 Searching the World Wide Web. *Science* **280**, 98–100.

Lawrence, S. & Giles, C. L. 1999 Accessibility of information on the web. *Nature* **400**, 107–109.

Newman, M. E. J. 2001 The structure of scientific collaboration-networks. *Proc. Natl Acad. Sci. USA* **98**, 404–409.

Nowak, M. A. 2000 The basic reproductive ratio of a word, the maximum size of a lexicon. *J. Theor. Biol.* **204**, 179–189.

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.

Nowak, M. A., Komarova, N. L. & Niyogi, P. 2001 Evolution of universal grammar. *Science* **404**, 114–118.

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

Simon, H. A. 1957 *Models of Man*. New York: Wiley.

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

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

Watts, D. J. 1999 *Small worlds: the dynamics of networks between order and randomness*. Princeton University Press.

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

Williams, R. J. & Martinez, N. D. 2000 Simple rules yield complex food webs. *Nature* **404**, 180–183.