Stability of Strength and Weight Distributions for Time-Evolving Word Co-occurrence Networks

Francesc Font-Clos and Álvaro Corral

1 Introduction

The most intriguing and celebrated empirical law in quantitative linguistics is Zipf's law [6], which in one of its forms states that the distribution of word frequencies in a text follows a power law with exponent $\gamma \sim 2$. At least in a qualitative sense, the fulfillment of Zipf's law is astonishing, being valid no matter the author, style, or language [4–6]. An important problem of Zipf's law is the variation of the exponent γ among different samples. Although the dependence of γ with system size was firstly acknowledged by Zipf himself [6], and later on other authors have confirmed it [1, 2], few systematic studies on these dependence have been performed. This can be formulated within the framework of (directed) networks, where words (types) are nodes, and consecutive appearances of word tokens increase the weight w_{ij} of a link between the two nodes by an amount equal to one. In this way, the frequency of a word is equivalent to the strength $s_i = \sum_j w_{ij}$ of its corresponding node.

2 Results

We propose a novel and simple scaling form for the strength distribution $P_N(s)$, Eq. (1), which uncovers the robustness of this distribution as the network size N is

F. Font-Clos (✉)
Departament de Matemàtiques, Universitat Autònoma de Barcelona, Barcelona, Catalonia, Spain

Centre de Recerca Matemàtica, Barcelona, Catalonia, Spain
e-mail: fontclos@crm.cat

Á. Corral
Centre de Recerca Matemàtica, Barcelona, Catalonia, Spain
e-mail: acorral@crm.cat

Á. Corral et al. (eds.), *Extended Abstracts Spring 2013*, Trends in Mathematics 2,
DOI 10.1007/978-3-319-08138-0_4

varied [3]. In this way, the shape of the distribution is always the same and it is only a scale parameter what increases linearly with system size,

$$P_N(s) = \frac{g(s/S)}{NS}, \qquad S = \sum_{i=1}^{N} s_i \,. \tag{1}$$

By analyzing the google's n-gram data set we verify this scaling and show that the strength distribution of the English language network has been extremely stable over the past 200 years; see Fig. 1. We also study the distribution of individual weights $P_N(\omega_{ij})$, reaching an analogous scaling form,

$$P_N(w) = \frac{h(w/S)}{WS}, \qquad W = \sum_{i,j=1}^{N} \delta(w_{ij}). \tag{2}$$

In addition, the growth of $P_N(s)$ and $P_N(w)$ with system size N allow us to estimate the number of nodes N and links W as a function of the total strength S,

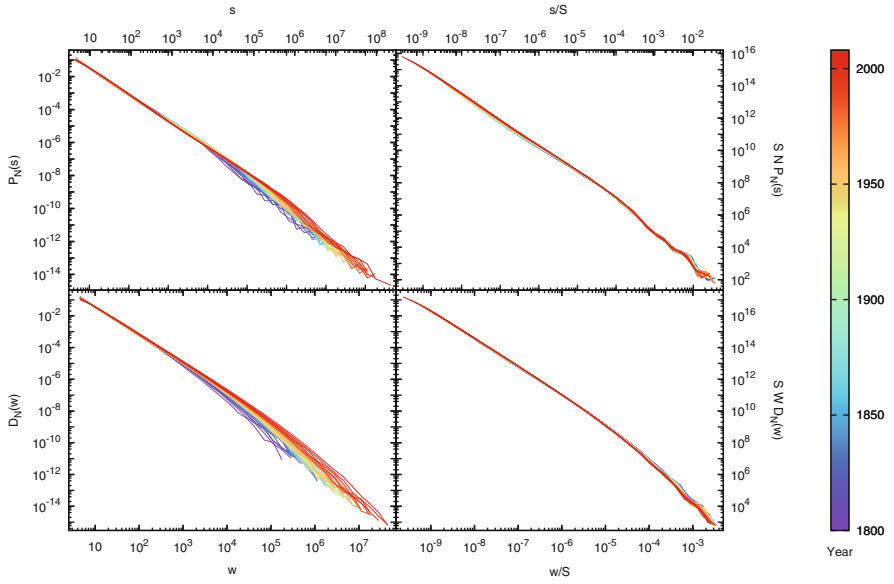

Fig. 1 $P_N(s)$ (*top*) and $D_N(w)$ (*bottom*) corresponding to the English language over the past 200 years before the rescaling process (*left*) and after it (*right*). The data collapse can be considered excellent, and the scaling functions $g(x)$ and $h(x)$ become apparent, uncovering the invariance over time of the English language network, both at the level of the strength distribution as well as at the level of the weight distribution

$$N(S) = \int_{1/S}^{\infty} g(x)\,dx, \tag{3}$$

$$W(S) = \int_{1/S}^{\infty} h(x)\,dx. \tag{4}$$

These equations provide a straightforward way to relate N and W, that is, the relation between the number of links and the number of nodes of language networks as time grows.

3 Discussion

Our findings suggest that the previous observation that the Zipf's exponent γ depends on system size [1, 2], might be an artifact of the increasing weight of a second regime in the strength distribution beyond a certain system size. The robustness of Zipf-like parameters under changes in system size opens the way to more practical applications of network science in linguistics. In particular, we provide a consistent way to compare statistical properties of language networks of different sizes.

References

1. H. Baayen, *Word Frequency Distributions* (Kluwer, Dordrecht, 2001)
2. S. Bernhardsson, L.E. Correa da Rocha, P. Minnhagen, The meta book and size-dependent properties of written language. New J. Phys. **11**, 123015 (2009)
3. F. Font-Clos, G. Boleda, A. Corral, A scaling law beyond Zipf's law and its relation with heaps' law. New J. Phys. **15**, 093033 (2013)
4. D. Zanette, Statistical patterns in written language (2012) http://fisica.cab.cnea.gov.ar/estadistica/zanette/papers/lang-patterns.pdf
5. D. Zanette, M. Montemurro, Dynamics of text generation with realistic zipf's distribution. J. Quant. Linguist. **12**(1), 29–40 (2005)
6. G.K. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley, Cambridge, 1949)