Givón, Talmy. 2000. 'Internal reconstruction: as method, as theory'. In Spike Gildea (ed.), *Reconstructing grammar: comparative linguistics and grammaticalization*, 107-159. Amsterdam: John Benjamins.

Heine, Bernd, Ulrike Claudi & Friederike Hünnemeyer. 1991. *Grammaticalization: a conceptual framework*. Chicago: University of Chicago Press.

Heine, Bernd & Mechthild Reh. 1984. *Grammaticalization and reanalysis in African languages*. Hamburg: Helmut Buske Verlag.

Hopper, Paul J. & Elizabeth Closs Traugott. 1993. *Grammaticalization*. Cambridge: Cambridge University Press.

Lehmann, Christian. 1983. 'Rektion und syntaktische Relationen'. *Folia Linguistica* 17, 339-378.

Lehmann, Christian. 1985. *Thoughts on grammaticalization*. München: Lincom Europa.

Lessau, Donald (ed.) 1994. *A Dictionary of grammaticalization*. 3 vols. Bochum: Universitätsverlag Dr. N. Brockmeyer.

Mikola, Tibor. 1975. *Die alten Postpositionen des Nenzischen (Jurak-samojedischen)*. Den Haag: Mouton.

Pinault, Georges-Jean 1989. 'Tokharien'. In *LALIES* 7, *Actes des sessions de linguistique et de littérature*. Paris: Presses de l'École Normale Supérieure, 6-224.

Lund University, Dept. of Linguistics and Phonetics
Working Papers 53 (2008), 15–21

Why is the Good distribution so good? Towards an explanation of word length regularity

Mats Eeg-Olofsson

Abstract
In 2004, Sigurd, Eeg-Olofsson & van de Weijer fitted the discrete analogue of the statistical gamma distribution to the frequency of the length of various linguistic units, in particular the length in letters of the word tokens in English and Swedish corpora. This distribution is also known as the Good distribution (Johnson, Kemp & Kotz 2005), named after I. J. Good, the statistician. In 2005, Lupsa & Lupsa successfully fitted this distribution also to the length of the base forms in Romanian and English dictionaries. Without further motivation, Lupsa & Lupsa call this regularity a linguistic law. This paper presents data from various languages to show that it is indeed a candidate for a linguistic universal and hints at some ways of explaining it.

Introduction

Sigurd, Eeg-Olofsson & van de Weijer 2004 investigated the word length distribution of the million-word corpora *Press-65* (Swedish) and *Brown Corpus* (American English), fitting it to the Good distribution. The Good distribution, which is a special case of the so-called Lerch distribution (Johnson, Kemp & Kotz 2005) is described by the formula:

$$f(l) = C \cdot l^a \cdot b^l,$$

where l is the length (in letters), $f(l)$ is the probability of length l, C a normalizing constant, and a and b parameters whose values depend on the particular language.

Fitting the Good distribution to more languages

For the work reported here, the *Regress+* software has been used to fit word type length data from six different European languages to the Good distribution. The data are based on the frequency word lists of the freely available *Leipzig corpora collection*, each sample containing about 100,000 sentences. The languages are English, Finnish, French, Sorbian (a Slavic

minority language), Swedish, and Turkish. The results are shown in Figures 1–6, where the *x* axis denotes word length in letters and the *y* axis word type frequency.

As can be seen from the graphs, the fit is very good. The goodness of fit can be measured by the R-squared statistic, which denotes the proportion of the variance in the data that is accounted for by the model, the Good distribution. In all cases, the *R*-square value exceeds 0.99, so the results gathered so far are encouraging.

Significance of the results

Thus, the Good distribution is a compact description of the statistical data. It can be used for checking automatically that language-like data are indeed natural language and not random noise, with potential applications in cryptology and speech understanding.

If the Good distribution of word type length is indeed a linguistic universal, it can also find an interesting (but somewhat esoteric) application in the search for messages from outer space. Data received from outer space are more likely to be intelligent signals if they exhibit some kind of word length distribution akin to that of the languages of the earth (Elliot, Atwell & Whyte 2000).

In search of explanation

It is hard to believe that the observed regularity is a mere coincidence. How can it be explained? Rather than delving into philosophical questions about the nature of explanation, I will suggest some lines along which it could be explained.

(a) Optimality – the distribution is optimal with respect to some kind of human cognitive or linguistic processing. In a certain sense, human languages are codes. Concepts from coding theory, like decodability and redundancy, might explain why human languages are optimal codes, possibly with respect to side conditions such as phonotactic constraints.

(b) Model genesis – the distribution is the outcome of some cognitive or linguistic process, possibly as an optimal or limiting distribution. Examples of such processes found in the literature are the 'naming games' described by Steels 1996.

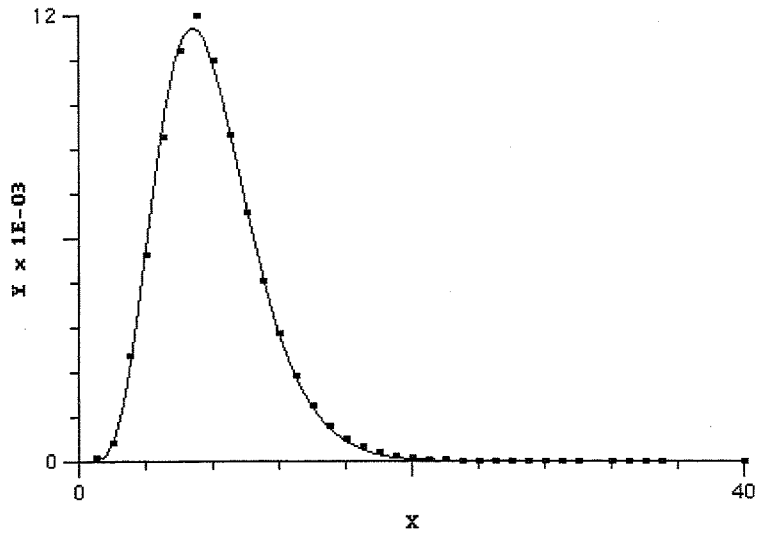

Figure 1. English ($a = 6.1, b = 0.41$).

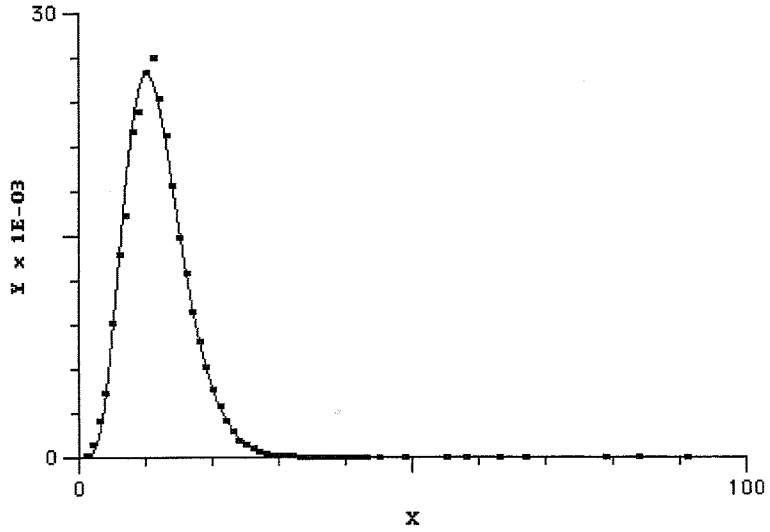

Figure 2. Finnish ($a = 5.9, b = 0.56$).

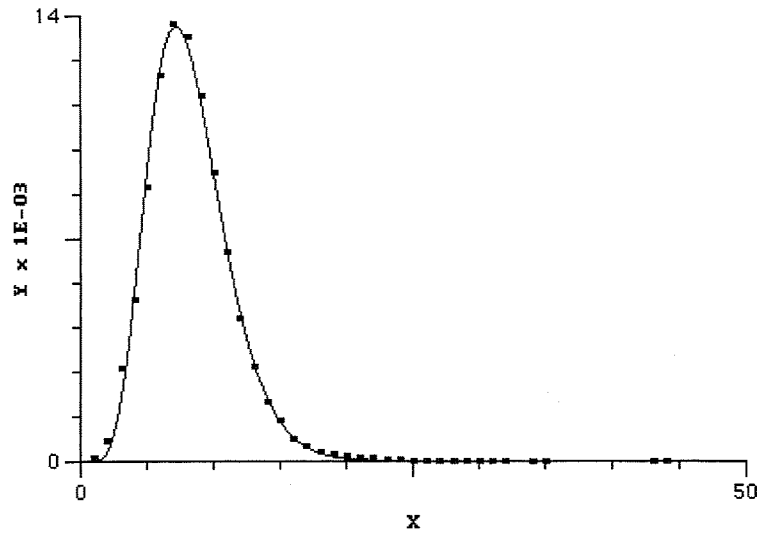

Figure 3. French ($a = 6.8, b = 0.39$).

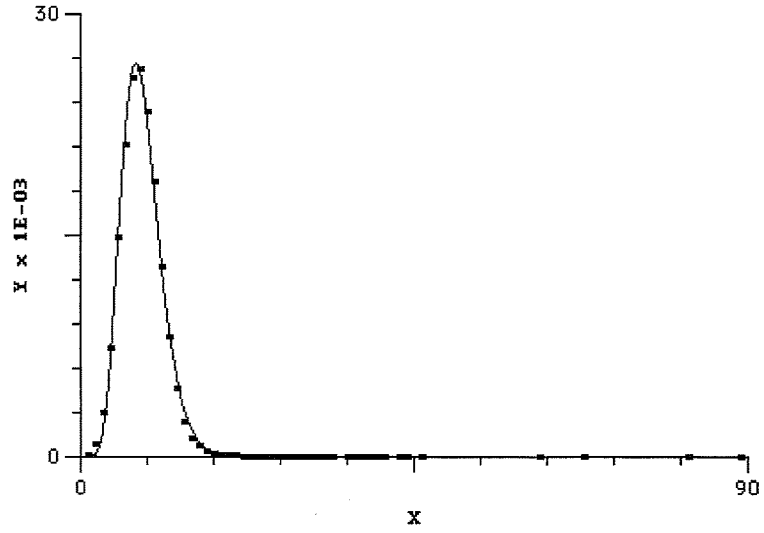

Figure 4. Sorbian ($a = 8.6, b = 0.32$).

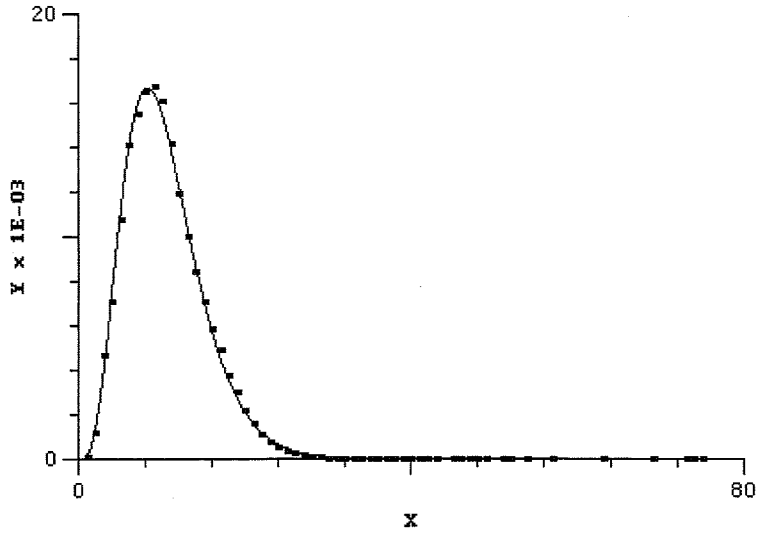

Figure 5. Swedish ($a = 3.9, b = 0.63$).

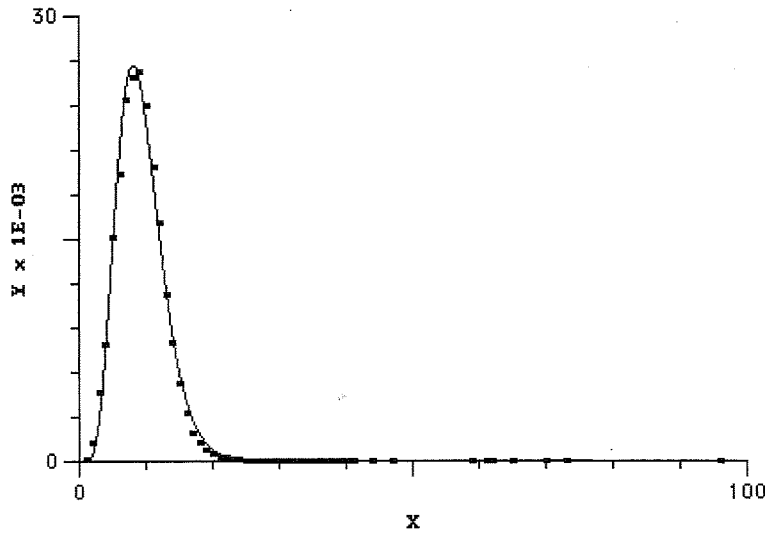

Figure 6. Turkish ($a = 6.3, b = 0.46$).

A tentative model genesis

Since the Good distribution is the discrete analogue of the gamma distribution, processes that give rise to the latter may be considered as explanations. Some such processes are the ones corresponding to a certain stochastic differential equation, whose stationary distribution is exactly the gamma distribution (Cobb 1998). The equation describes the length distribution as a result of both systematic and random change during the historical development of language. This equation is:

$$dx = r(G-x)dt + \sqrt{\varepsilon x}dw$$

where $x = x(t)$ is the length of the expression of a certain concept at time t and G is a 'goal' value for length. The term $r(G-x)dt$ on the right hand side of the equation expresses that the length process is subject to linear feedback with respect to the goal value G. The second term on the right hand side expresses that the diffusion (fluctuation) $\varepsilon x(t)$ at time t is proportional by the constant ε to the length $x(t)$ at time t.

Thus, the critical assumptions of this model are the following:

(a) The length of a word is attracted (by linear feedback) to a goal value G.

(b) Long words (i.e. long expressions of concepts) are subject to greater 'disturbance' than short words in the course of language development.

It would be interesting to find empirical evidence for these assumptions in the history of language.

Conclusion

The results seem to express a generalization over typologically and genetically distinct languages. Thus the Good distribution for lexical word length is a candidate for a linguistic universal. It would be interesting to strengthen the evidence by obtaining data from additional languages, including other alphabetic writing systems.

Acknowledgements

I am indebted to Bengt Sigurd for valuable comments on an earlier version of this paper.

References

Cobb, Loren. 1998. 'Stochastic differential equations for the social sciences'. Revised and extended from Chapter 2 of Cobb & Thrall (eds.), *Mathematical frontiers of the social and policy sciences*. Westview Press, 1981. Website: http://www.aetheling.com/docs/SDE.pdf accessed 2007-03-11.

Elliot, John, Eric Atwell & Bill Whyte. 2000. 'Increasing our ignorance of language: identifying language structure in an unknown signal'. *Proceedings of the Fourth Conference on Computational Natural Language Learning and of the Second Learning Language in Logic Workshop*, 25-30. Lisbon.

Johnson, Norman L, Adrienne W. Kemp & Samuel Kotz. 2005. *Univariate discrete distributions*. Third edition. Hoboken, NJ: John Wiley.

Leipzig Corpora Collection. Website: http://corpora.informatik.uni-leipzig.de/download.html accessed 2007-03-11.

Lupsa, Dana Avram & Radu Lupsa. 2005. 'The law of word length in a vocabulary'. *Studia univ. Babes-Bolyai, Informatica* 50:2, 69-80.

Regress+. Website: http://www.causascientia.org/software/Regress_plus.html accessed 2007-03-11

Sigurd, Bengt, Mats Eeg-Olofsson & Joost van de Weijer. 2004. 'Word length, sentence length and frequency (Zipf revisited)'. *Studia Linguistica* 58, 37-52.

Steels, Luc. 1996. 'Self-organizing vocabularies'. In Christopher G. Langton & Katsunori Shimohara (eds.), *Artificial Life* 5, 179-184. Nara, Japan.