On the intermittent dyanmics of language use

Eduardo G. Altmann,[1, 2, 3, *] Janet B. Pierrehumbert,[3, 4] and Adilson E. Motter[3, 5]

[1] *Instituto de Física, Universidade Federal do Rio Grande do Sul, 91501-970 Porto Alegre, Brazil*
[2] *Max Planck Institute for the Physics of Complex Systems, 01187 Dresden, Germany*
[3] *Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL 60208, USA*
[4] *Department of Linguistics, Northwestern University, Evanston, IL 60208, USA*
[5] *Department of Physics and Astronomy, Northwestern University, Evanston, IL 60208, USA*

This talk will report a detailed statistical analysis of the temporal distribution of words over times much longer than sentence length. Inspiration for our work [1] comes from Zipf's discovery [2, 3] that word frequency distributions obey a power law. This discovery established parallels between biological and physical processes, and language, laying the groundwork for a complex systems perspective on human communication. More recent research has also identified scaling regularities in the dynamics underlying the successive occurrences of events, suggesting the possibility of similar findings for language as well. By considering frequent words in USENET discussion groups and in disparate databases where the language has different levels of formality, we investigate the distribution $f(\tau)$ of distances τ between successive occurrences of the same word. We show that words typically display bursty deviations from a Poisson process

$$f_P(\tau) = \mu e^{-\mu\tau}, \tag{1}$$

and are well characterized by a stretched exponential (Weibull) scaling:

$$f_\beta(\tau) = a\beta\tau^{\beta-1}e^{-a\tau^\beta}, \tag{2}$$

The extent of this deviation is quantified by β, a measure of the burstiness of each word (for different quantifications of this behaviour see Refs. [4–7]). The observed distances are much longer than sentence length indicating the connection between the burstiness and the semantics of the words. To investigate this connection we classify all words according to their semantic type – a measure of the logicality of each word explained in Tab. I. Figure 1 summarizes a detailed analysis over more than two thousand frequent words. Values of β close to 1 indicate words closer to the random (Poisson) process, while smaller values of β indicate a strong bursty behaviour. Higher Class words tend to have β close to 1, while low Class words tend to have small values of β. The comparison in the right panel of Fig. 1 shows that β depends more strongly on semantic type than on frequency $\nu = 1/\langle\tau\rangle$. We develop a generative model that fully determines the dynamics of word usage. Because the use of words provides a uniquely precise and powerful lens on human thought and activity, our findings also have implications for other overt manifestations of collective human dynamics.

Class	**Name**	**Examples of words**
1	Entities	Africa, Bible, Darwin
2	Predicates and Relations	blue, die, in, religion
3	Modifiers and Operators	believe, everyone, forty
4	Higher Level Operators	hence, let, supposedly, the

TABLE I: Examples of the classification of words by semantic types. The primitive types are entities e, exemplified by proper nouns such as *Darwin* (Class 1), and truth values, t (which are the values of sentences). Predicates or relations, such as the simple verb *die*, and the adjective/noun blue, take entities as arguments and map them to sentences (e.g., *Darwin dies*, *Tahoe is blue*). They are classified as $< e, t >$ (Class 2). The notation $< x, y >$ denotes a mapping from an element x in the domain to the image y [8, 9]. The semantic types of higher Classes are established by assessing what mappings they perform when they are instantiated. For example, *everyone* is of type $<< e, t >, t >$ (Class 3), because it is a mapping from sets of properties of entities to truth values [9]; the verb *believe* shares this classification as a verb involving mental representation. The adverb *supposedly* is a higher order operator (Class 4), because it modifies other modifiers. Following Ref. [9] (contra Ref. [8]) words are coded by the lowest type in which they commonly occur (see Text S1, *Coding of Semantic Types*).

* Electronic address: egaltmann@gmail.com

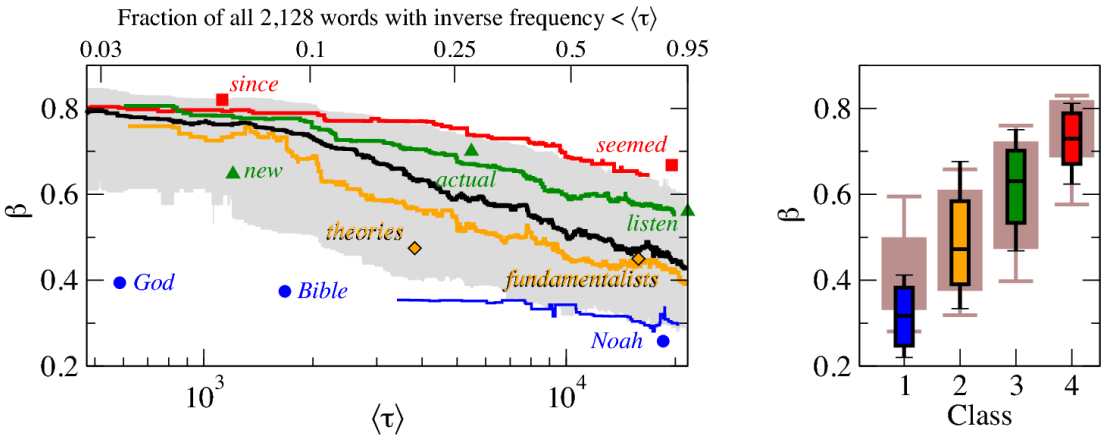

FIG. 1: (left) Relative dependence of β on Class and $\langle\tau\rangle = 1/\nu$ (inverse frequency), indicating: running median on words ordered according to $\langle\tau\rangle$ (center black line) and 1-st and 7-th octiles (boundaries of the gray region); and running medians on words by Class (colored lines, Class 1-4, from bottom to top) with illustrative words for each Class. At each $\langle\tau\rangle$, large variability in β and a systematic ordering by Class is observed. (right) Box-plots of the variation of β for words in a given Class. The box-plots in the background are obtained using frequency to divide all words in four groups with the same number of words of the semantic Classes (first box-plot has words with lowest frequency and last box-plot has words with highest frequency). The classification based on Classes leads to a narrower distribution of β's inside Class and to a better discrimination between Classes.

[1] Altmann EG, Pierrehumbert JB, and Motter AE (2009) Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. PLoS ONE 4 (11) e7678.
[2] Zipf GK (1935) The Psycho-biology of Language: An Introduction to Dynamic Philology. Boston: Houghton Mifflin.
[3] Zipf GK (1949) Human Behavior and the Principle of Least Effort. New York: Addison-Wesley.
[4] Church KW, Gale WA (1995) Poisson mixtures. Nat Lang Eng 1:163–190.
[5] Katz SM (1996) Distribution of content words and phrases in text and language modelling. Nat Lang Eng 2:15–59.
[6] Montemurro MA, Zanette DH (2002) Entropic analysis of the role of words in literary texts. Advances in Complex Systems 5:7–17.
[7] Ortuño M, Carpena P, Beranaola-Galván P, Muñoz E, Somoza AM (2002) Keyword detection in natural languages and DNA. Europhys Lett 57:759–764.
[8] Montague R (1973) The proper treatment of quantification in ordinary English. In: Hintikka J, Moravscik J, Suppes J, editors. Approaches to Natural Language. Dordrecht: Reidel. pp. 373–398.
[9] Partee BH (1992) Syntactic categories and semantic type. In: Rosner M, Johnson R, editors. Computational Linguistics and Formal Semantics. Cambridge: Cambridge Univ. Press. pp. 97–126.