Towards a Stochastic Model of the Human Word-finding Process Underlying Zipf's Law: A Crucial Role for Sample-Space Reduction?

Gerard Kempen (MPI for Psycholinguistics, Nijmegen, The Netherlands)
Karin Harbusch (Faculty of Computer Science, University of Koblenz, Germany)
Gerard.Kempen@MPI.NL

We propose a "bounded rationality" model of the emergence of Zipf's Law in word frequency distributions. It assumes that *Sample-Space Reduction* (SSR) as defined by Corominas-Murtra et al. (2015/16) and Thurner et al. (2015/18; henceforth CH&T) can model a key phenomenon of human language production: semantic precision being compromised in favor of easier lexical access. Zipf (1936/1949) himself conjectured a causal link between this "least effort" tendency and the frequency distributions he had observed: power-law distributions with slope parameter $\alpha \approx 1$ (Fig. 1). However, no-one has since proposed a cognitively plausible theory of why "least effort" yields distributions close to Zipf's Law (see review by Piantadosi 2014).

Selection of lexical items during language production is standardly depicted as a three-stage process: from reference delimitation via concept activation to lemma selection. (Lemmas correspond to citation forms of inflected wordforms.) In line with lexicographic practice, we assume that many concepts (meanings) are associated with one or more (synonymous) lemmas, and that concepts vary w.r.t. semantic complexity: the number of criteria determining whether the activated concept accurately covers the intended reference (denotation)—neither too broad nor too narrow. If such a lemma proves hard to access, producers will resort to *referentially "good enough" concepts* associated with more easily accessible lemmas. Options include (1) switching to a concept that delimits the reference by applying another set of criteria; (2) selecting a superordinate concept (a simpler, less precise meaning), and/or (3) splitting the delimitation criteria across multiple, simpler concepts and conceptual dependency links (thereby often restoring semantic precision). Crucially, these scenarios cause a *unidirectional frequency shift*: it boosts the frequencies of lemmas with relatively imprecise meanings, and of "function words" (many of them used to mark conceptual dependencies explicitly). This contributes to a negative correlation between the referential precision and the usage frequency of content and function lemmas.

The "good enough" ("satisficing") word-finding strategy generates ranked lemma-frequency distributions with heads densely populated by a small vocabulary of semantically imprecise but easily accessible content and function words, and with tails sparsely populated by a large set of more precise but harder to find lemmas. CH&T present mathematical proof and computer simulations of a remarkable result: SSR transforms large, relatively flat *input* power laws ($0 \leq \alpha' < 1$) into *output* power laws with $\alpha \approx 1$. This outcome obtains ("in the limit", and in the absence of external biases) with input distributions spanning large (including human) vocabularies, generalizing beyond power laws to many types of frequency distributions with zero or negatively accelerated decay. In the words of CH&T: *Zipf's Law acts as an attractor* (Fig. 2).

We propose to treat concept-frequency and lemma-frequency distributions as input and output distributions, respectively, hypothesizing that "good enough, easy-access" word finding tendencies will map the former onto the latter by emulating SSR. This presupposes that slope exponents of ranked concept-frequency distributions do not exceed 1. If this can be verified, and if additionally observed details of human word-finding turn out compatible with the assumptions underlying SSR, the proposed model will meet an important criterion put forward in Piantadosi's (2014) review: that any explanation of Zipf's Law should be founded on a plausible view of lexical processing.

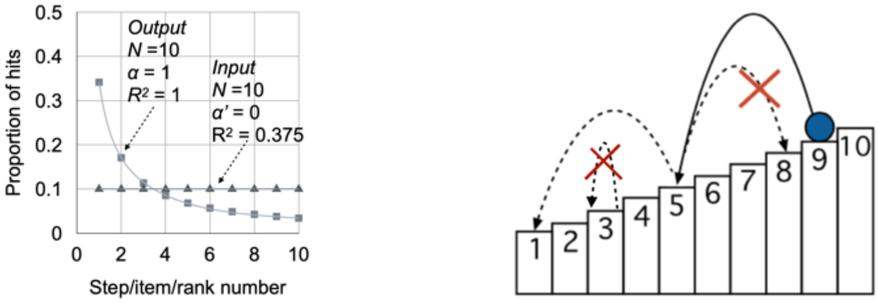

Fig. 1. “Zipf’s Law” emerging from a uniform “input” distribution. *LEFT*: A “power law” is a ranked distribution of item probabilities in which the probabilities of rank *i* are proportional to those of a harmonic series: $p(r_i) = 1/i$ for $i = 1, \ldots, N$); e.g., the curve labeled “output”. The slope of power-law curves can be adjusted by raising the denominators to a power $\boldsymbol{\alpha}$; $p(r_i) = 1/r_i^{\alpha}$. For $\boldsymbol{\alpha} > 1$, decay is steeper, for $\alpha < 1$ it is flatter than that of a power law with $\boldsymbol{\alpha} = 1$, i.e., the slope of “Zipf’s Law” proper. *RIGHT*: This stairway (drawing slightly adapted from CH&T) illustrates the notion of *Sample-Space Reduction*. Imagine a ball is bouncing down the steps, never rebounding to a higher step (unidirectionality), hitting (“visiting”, “sampling”) the same step at most once, and halting at the lowest step. At the onset of each jump, the ball has a number of contiguous steps to chose from: the current “sample space”. The probability of the ball visiting r_i during a jump equals 1 divided by the current sample space. This yields a harmonic series if the steps have equal widths, hence equal probabilities of being sampled. We refer to a distribution of step widths as “input distribution”. In the left chart, the input distribution is *uniform* (which, analyzed as power law means $\alpha' = 0$, with low R^2). However, the steps may have wider and narrower widths, causing them to be visited with proportionately higher or lower probabilities (discussed in Fig. 2).

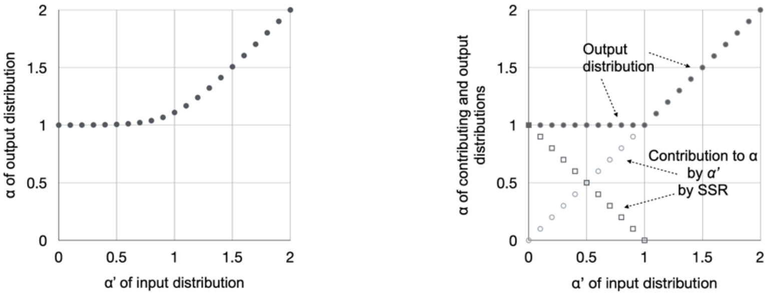

Fig. 2. Effects of applying SSR to input distributions decaying with varying slopes (*α*’). *LEFT*: SSR applied to *input* distributions with slope $0 \leq \alpha' \leq 2$, spanning $N = 50{,}000$ steps. This *N* value reflects common estimates of the active lemma vocabularies of adult natural-language users. The width distributions of the steps are power laws with either a flat input distribution ($\alpha' = 0$), a slow decay rate ($0 < \alpha' \leq 1$), or a rapid decay rate ($\alpha' > 1$). SSR tends to cause accumulation of probability mass at the head of the output distribution, thereby attenuating the probability mass occupied by the tail. With large and slowly decaying input distributions, the emerging output slope values remain within a very narrow bandwidth around $\alpha \approx 1$: “Zipf’s Law as an attractor.” *RIGHT*: These nearly invariant output slope values can be understood intuitively as *additive* contributions of *α*’ and SSR to the slope of output distributions. The chart represents the situation expected when *N* is approaching infinity. Open and filled circles: contribution by *α*’; open squares: contribution by SSR; filled circles: slope values of the emerging output distributions. For instance, at $\alpha' = 0$ (uniform, horizontal input distribution), SSR is responsible for the entire output slope ($\alpha = 1$), yielding a harmonic series. For larger values of *α*’, the SSR contributions decrease: a higher *α*’ implies a thinner tail, hence lower probabilities of downward jumps from high-rank steps belonging to the tail. SSR runs dry at $\alpha' = 1$, meaning $\alpha = \alpha'$ for $\alpha' \geq 1$. For the proof see CH&T.

References

Corominas-Murtra, B., Hanel, R., & Thurner, S. (2015). Understanding scaling through history-dependent processes with collapsing sample space. *PNAS, 112,* 5348-5354.

Corominas-Murtra, B., Hanel, R., & Thurner, S. (2016). Extreme robustness of scaling in sample space reducing processes explains Zipf’s law in diffusion on directed networks. *New J. Phys., 18*, 093010.

Piantadosi, S. (2014). Zipf’s word frequency law in natural language: A critical review and future directions. *Psychon. Bull. Rev., 21*, 1112-1130.

Thurner, S., Hanel, R., Liu B., & Corominas-Murtra, B. (2015). Understanding Zipf’s law of word frequencies through sample-space collapse in sentence formation. *J. R. Soc. Interface, 12*, 20150330.

Thurner, S., Hanel, R., & Klimek, P. (2018). *Introduction to the Theory of Complex Systems.* OUP.

Zipf, G.K. (1936). *The psycho-biology of language*. Routledge.

Zipf, G.K. (1949). *Human behavior and the principle of least effort.* AddisonWesley.