Historical, Lifespan, and Individual Variation in the Zipf Distribution of Prolific Novelists

Geoff Hollis & Chris Westbury
University of Alberta

George Kinsey Zipf observed that, within a given language, there exists a log-linear relation between the probability that a word will occur and the word's rank position in the ordered list of words sorted by descending frequency (Zipf, 1935). This relation is regular and generalizes across languages, even with different etymological roots. Zipf originally framed the phenomenon as an emergent property that falls directly out of the divergent requirements of senders and receivers. Within the context of communication, a sender's effort is minimized by reducing the number of word forms used for communication; *communication efficiency* is at its highest when a lexicon contains a single word that is used to convey all meanings. In comparison, a receiver's effort is minimized by increasing the number of word forms in a lexicon; *communication effectiveness* is at its highest when each possible meaning conveyed has its own distinct word form. The interacting demands of receivers and senders give rise to a lexicon with the power law distribution, $p(\mathrm{i}) \sim \mathrm{i}^{-a}$ where i is the frequency of *i*-th most-frequent word, and *p*(i) is the probability of encountering that word. Changes in the exponent within a single lexicon and generating system are interpretable in information theoretic terms, with increasing magnitude of the exponent entailing increased Shannon information (more distinctions) in the communicative content.

Natural Languages invariably have an a close to 1. This is known as Zipf's Law. However, the particular power that best characterizes a non-arbitrary subgroup's linguistic productions shows substantial variability. In our research, we have examined how the power law exponent of a language changes across historical time as well as within the lifespan of individual language users. Our measurements are based on a sample of 690 book-length texts written by prolific authors with entries within the Project Gutenberg database. We have documented that there is an historical trend for a decrease in the magnitude of the exponent across history, with a correlation between date of publication and the power law exponent of 0.28 (df = 645, df < 0.001). Since communication with a lower magnitude exponent suggests increased communication efficiency (easier for the producer of communication) but decreased communication effectiveness (harder for the receiver of communication), this suggests that author constraints have played an increasing role in shaping written output across time.

The exponent characterizing an author's work varies within their lifespan, in a non-linear U-shaped curve described by an equation whose first and second order regression terms for age are reliably different than zero (t[58] >= 2.3, p < 0.05 in both cases). The magnitude of the exponent for an author's works tends to increase until about the age of 55 and then decrease. We consider two hypotheses for explaining this finding: that it may reflect degeneration in acquired knowledge (older authors are less able to maximize communicative efficiency for readers) or that authors may have less concern later in their careers for maximizing communicative efficiency for their readers. Since the first hypothesis is age-based and the second career-length based, it is possible to adjudicate between these possibilities. There is a negative correlation between career length and the magnitude of the distribution exponent (rather than the same U-shaped curve we see when we look at age), which gives support to the biologically-based explanation.

To our knowledge, this is the first longitudinal study of how the distributional characteristics of communication vary across time.

References

Zipf, G. K. (1935/1965). *The Psycho-biology of Language: An introduction to Dynamic Philology.* Cambridge, MA: MIT Press.