Towards a Rigorous Motivation for Zipf's Law

Phillip M. Alday

University of South Australia

Cognitive Neuroscience Laboratory

Introduction

Models of language fall broadly into two categories:

Models of the communicative system are often formulated as mathematical models based on simple distributional properties of language, commonly presented as empirical laws.

Psychological and neurobiological models have focused largely on the computational constraints presented by incremental, real-time processing.

Information-theoretic entropy underpins successful models of both types and provides a more principled motivation for Zipf's Law.

Zipf's Law for frequency

Zipf [1–3] demonstrated that distributional statistics in language often follow a power law. In particular, the relationship between frequency (f) and rank (r) is given by:

$$f \propto \frac{1}{r} \Leftrightarrow f = \frac{c}{r} \qquad \text{for some constant } c \tag{1}$$

This is often extended via an exponent, empirically observed to be near 1, allowing for an easily estimated slope parameter when plotted log-log:

$$f = \frac{c}{r^\alpha} \Rightarrow \log f = \log \frac{c}{r^\alpha} = \log c - \alpha \log r \tag{2}$$

Ideal relationship between c and α

The probability density function (PDF) for the Pareto distribution is given by

$$P(x) = \frac{(\alpha - 1)x_0^{\alpha-1}}{x^\alpha}, x \geq x_0 \tag{3}$$

For rank data, we know that $x = r \geq 1$, which yields

$$P(r) = \frac{\alpha - 1}{r^\alpha}, \quad x \geq 1 \tag{4}$$

We recognize this as Zipf's Law, when $c = \alpha - 1$.

The emergence of Zipfian distributions in communication

- Zipf motivated this power law by his *principle of least effort*, but did not provide a rigorous motivation or description of this principle, i.e. how and why these frequency distributions came to be.
- More recent accounts show that Zipf's Law emerges naturally from a minimal set of assumptions about combined speaker-hearer effort.[4]
- Several empirical estimates have shown that α tends to be close to one.[5] Moreover, during first language acquisition $\alpha \to 1$, at least in the Germanic languages.
- Nonetheless, the exponent is often viewed as a free parameter.[5]

Linking Brains and Behavior: Words as Experiments

- Friston proposed a theory of neurocomputation based on the fitting of generative models of upcoming perceptual stimuli via expectation maximization.[6, 7]
- Behavior is part of the model-fitting process, determining how new data is sampled.[8]
- An accurate model follows from minimizing the (information-theoretic) free energy and surprisal in the generative models.
- This is achieved by seeking out the most surprising – i.e. informative – stimuli.
- Language should maximize average surprisal.

Information-theoretic Entropy

Using information theory, we can define *surprisal* or self-information rigorously as

$$I(x) = -\log P(x) \tag{5}$$

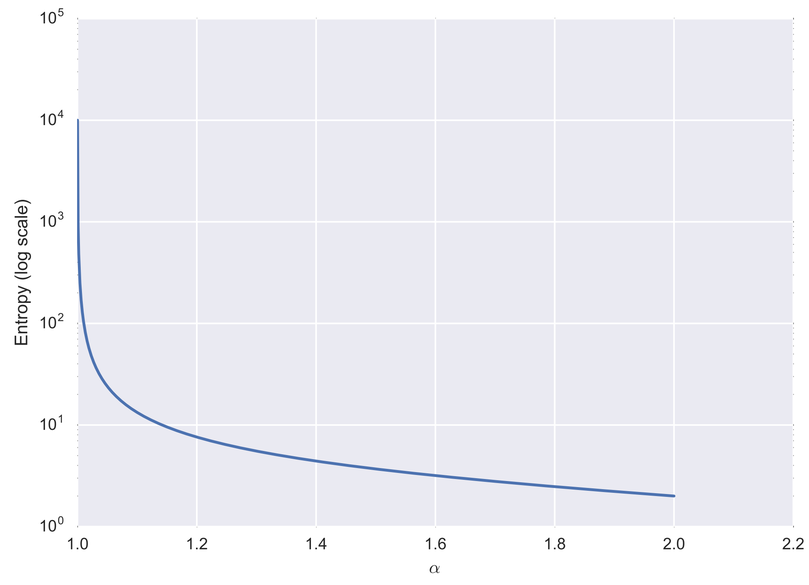
The logarithmic transform provides power-law type scaling. Crucially, the less probable a certain element is, the closer its probability is to zero and hence the further its logarithm is away from zero, i.e. the greater its surprisal.

The expected value of surprisal across an entire set is called *entropy* and is given by

$$H(X) = -\int_{x \in X} P(x) \log P(x)\, dx \tag{6}$$

Maximizing entropy in language: tuning the free parameters

From the Pareto PDF (4):



$$H(X) = \log\left(\frac{1}{\alpha - 1}\right) + \left(\frac{\alpha}{\alpha - 1}\right) \tag{7}$$

As $\alpha \to \infty$, $P(x)$ converges to the Dirac delta-function $\delta_{x_0=1}(x)$ and entropy drops as only one symbol (word) from a large pool is meaningful.

As $\alpha \to 1$, the distribution becomes successively flatter, but maintaining a spike-like structure with a thick tail.

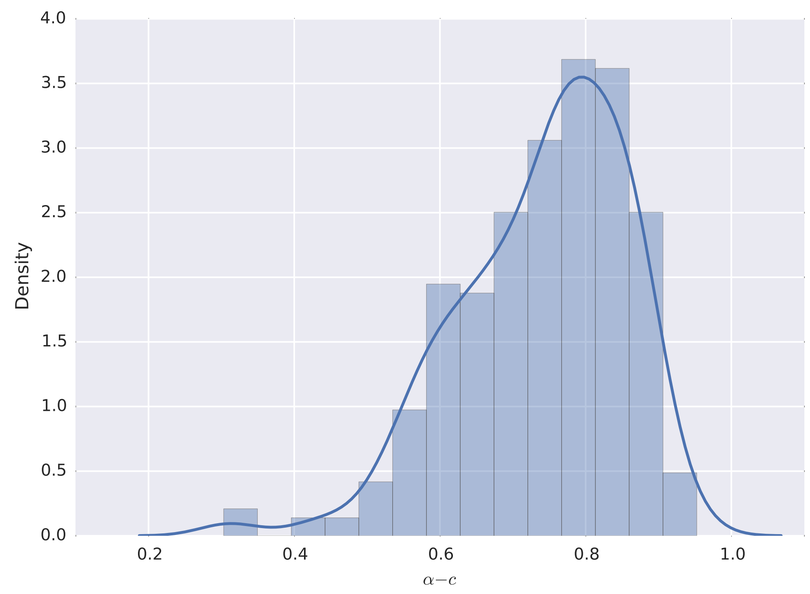
As such, we expect that $\alpha = 1$ is near optimal when $c = \alpha - 1$ and that languages will have evolved to have near optimal α.

Empirical basis

- 310 languages using the translations for the Universal Declaration of Human Rights provided by the `nltk.corpus` Python package.[9]
- OLS regression provides estimates for the intercept ($\log c$) and slope ($-\alpha$) from Equation (2).

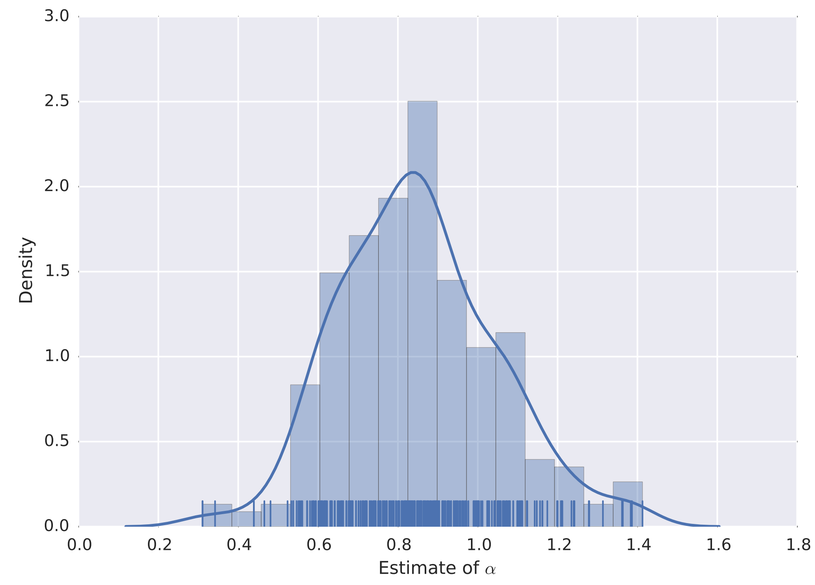
Actual relationship between α and c

Our ideal estimate assumes $c = \alpha - 1$.



But c is closer to $\alpha - 0.8$ which will bias empirical α.

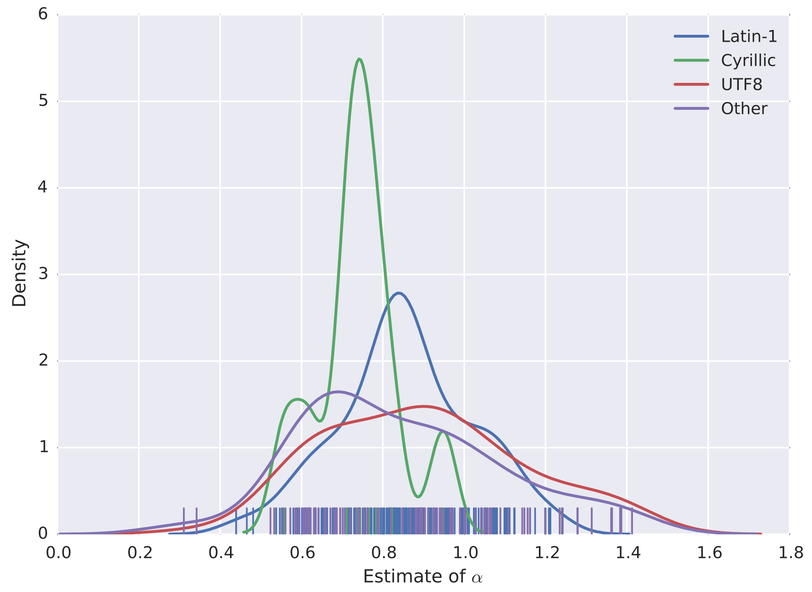
Empirical estimates across languages



Across languages, α has a mean of about 0.9. This is somewhat less than the original predicted idealized value of $\alpha = 1$; however, it is line with the bias we observed in c.

Variation between scripts

Because we are using an orthographic notion of word, we can consider the influence of the type of orthography on these estimates.



Encoding	n	α
Latin1	190	0.86 ± 0.01
Cyrillic	10	0.74 ± 0.03
UTF8	86	0.90 ± 0.03
Other	110	0.85 ± 0.02
All	310	0.85 ± 0.01

Conclusion

Frequency-based explanations beg the question *X does this because X is more frequent, but how did X become more frequent in the first place?*

- We have provided a principled motivation and empirical validation for the free parameter α.
- The observed bias in c suggests why previous work has found α to be near one, but rarely exactly one, even when corrected for observation error.
- Together with previous work, this provides a principled *causal* explanation for the emergence of Zipfian frequency distributions.
- This is a first step towards grounding empirical laws in the processing constraints and strategies of individual language users.

Having parameters that relate back to assumptions about basic cognitive strategies and processing constraints is far more valuable than having parameters related to uninformed curve fitting.

Literature

[1] G. K. Zipf (1929). *Harvard Studies in Classical Philology*.
[2] G. K. Zipf (1935). *The Psycho-Biology of Language*. Boston, MA: Houghton Mifflin Company.
[3] G. K. Zipf (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley.
[4] R. Ferrer-i-Cancho & R. V. Solé (2003). *Proceedings of the National Academy of Sciences*.
[5] I. Moreno-Sánchez, F. Font-Clos, et al. (2016). *PLoS ONE*.
[6] K. Friston (2005). *Philosophical Transactions of the Royal Society B: Biological Sciences*.
[7] K. Friston (2009). *Trends in Cognitive Sciences*.
[8] K. Friston, R. Adams, et al. (2012). *Frontiers in Psychology*.
[9] S. Bird, E. Klein, et al. (2009). *Natural Language Processing with Python*. O'Reily Media.