TOWARDS A RIGOROUS MOTIVATION FOR ZIPF'S LAW

PHILLIP M. ALDAY

Cognitive Neuroscience Laboratory, University of South Australia
Adelaide, Australia
phillip.alday@unisa.edu.au

Language evolution can be viewed from two viewpoints: the development of a communicative system and the biological adaptations necessary for producing and perceiving said system. The communicative-system vantage point has enjoyed a wealth of mathematical models based on simple distributional properties of language, often formulated as empirical laws. However, beyond vague psychological notions of "least effort", no principled explanation has been proposed for the existence and success of such laws. Meanwhile, psychological and neurobiological models have focused largely on the computational constraints presented by incremental, real-time processing. In the following, we show that information-theoretic entropy underpins successful models of both types and provides a more principled motivation for Zipf's Law.

1. Introduction

There are two distinct developments that the "evolution of language" refers to, namely (1) the biological, and especially the neurobiological, adaptations necessary for producing, perceiving and processing language and (2) the development of the communication system, abstracted away from generators and receivers of the signal. The development of the communication system has proven remarkably easy to study using rather simple mathematical models (e.g. Ferrer-i-Cancho, 2015b; Lieberman, Michel, Jackson, Tang, & Nowak, 2007; Pagel, Atkinson, & Meade, 2007; Piantadosi, Tily, & Gibson, 2011), but finding an equally parsimonious quantitative model for the neurobiology of language has proven surprisingly difficult. Despite advances on both sides, a combined approach has not been widely adopted, with the mathematical community at times dismissing the "psychological bias", much less the neuroscientific one (cf. Ferrer-i-Cancho, 2015a). A current neurobiological theory for cortical responses (Friston, 2005) provides the necessary unifying perspective for the evolution of language both as a communication system and as a neurobiological system. More precisely, the diachronic development is shaped by the synchronic constraints arising from basic neurocomputational principles. In the following, we will use this to derive Zipf's Law in the abstract from neurobiologically motivated first principles as well as provide a prediction about the form of its parameters.

2. Zipf's Law and Psychological Tradeoffs in Communication

Frequency-based explanations are common in empirical linguistics, yet they provide few deep, causal explanations (e.g. what drives the development of the frequency distribution?). Nonetheless, distributional statistics provide a convenient, largely theory agnostic method for modelling properties of a language. Zipf (1929, 1935, 1949) demonstrated that distributional statistics in language often follow a Pareto-like distribution (although that is not the terminology he used). Zipf suggested a number of power laws in language, but in the following we will focus on the relationship between frequency (f) and rank (r):

$$f \propto \frac{1}{r} \Leftrightarrow f = \frac{c}{r} \tag{1}$$

for some constant c. This is often extended via an exponent, empirically observed to be near 1, allowing for a slope parameter when plotted log-log:

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

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

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

where $x_0 > 0$ is the location parameter and expresses the minimum possible value and $\alpha > 1$ is the shape parameter[a] and expresses how "bent" the distribution is. When x's are ranks, then $x_0 = 1$ and this reduces to

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

which we recognize as a special case of Equation (2) when $c = \alpha - 1$.

Zipf postulated a principle of least effort as the motivation for his empirical laws, and indeed this matches well with the "80-20" laws often associated with the Pareto distribution. Ferrer-i-Cancho and Solé (2003) added mathematical rigor to this intuition via simultaneous optimization of hearer and speaker effort when operating on signal-object associations and showed that Zipfian distributions emerge naturally when hearer and speaker effort are weighted equally.

3. Linking Brains and Behavior: Words as Experiments

Friston (2005, 2009) proposed a theory of neurocomputation based on the fitting of generative models of upcoming perceptual stimuli via expectation maximization. Friston, Adams, Perrinet, and Breakspear (2012) expanded upon this proposal by incorporating action into the model-fitting process, focusing on saccades

[a]Traditionally, the PDF is expressed with α and $\alpha + 1$ such that $\alpha > 0$, but our presentation makes the notation more compatible with the literature on Zipf's Law.

(eye movements) in visual processing. An accurate model follows from minimizing the (information-theoretic) free energy and surprisal in the generative models. However, in order to best improve the generative model, the most informative, and therefore the *most surprising* stimuli, are sought out.

3.1. *Information-theoretic Surprisal*

In information-theoretic terms, *surprisal* is also called self-information and is defined as

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

i.e. the self-information of a specific element, class or form is the negative logarithm of the probability of its occurrence. The logarithmic transform provides power-law type scaling and turns additive effects on this scale into multiplicative effects on the original scale. Because probabilities are always between zero and one (inclusive), the logarithm is always negative and thus the negative sign in the definition places self-information on a non-negative scale. Although "self-information" and "surprisal" are technical terms with a precise definition, they nonetheless correspond roughly to intuition. The less probable a certain element is (i.e. the less expected it is), the closer its probability is to zero and hence the further its logarithm is away from zero, i.e. the greater its surprisal. Moreover, they contain more information in themselves because they are not as easily predictable.

3.2. *Information-theoretic Entropy*

We can also consider the amount of information contained in an entire set, or, equivalently, how much surprisal we should expect from a "typical" or "average" element. In technical terms, the expected value is given by:

$$H(X) = E[I(x)] = -\int P(x) \log P(x)\, dx, \quad x \in X \tag{6}$$

This value is commonly called *entropy*.

3.3. *Maximizing Entropy in Language*

If we assume that language is optimized for the balance between hearer and speaker, then we can replace $P(x)$ by the Pareto PDF (3,4) above and can maximize the entropy of language, i.e. the average surprisal, by optimizing the parameter α.

In particular, the entropy of the Pareto distribution (with $x_0 = 1$) is given by:[b]

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

[b]The derivation of this result is beyond the scope of this paper. Again, we use a slightly non traditional parameterization to better match the literature on Zipf's Law.

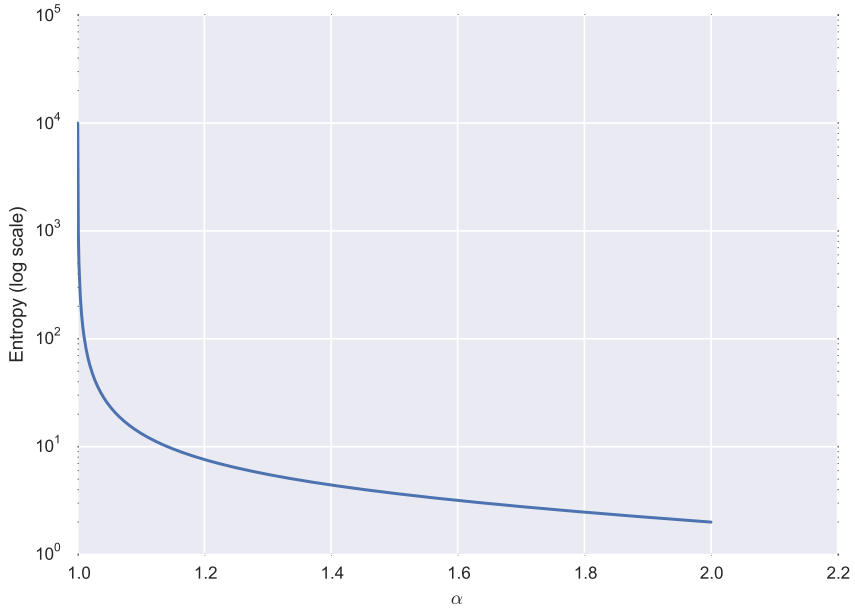

Figure 1. The entropy of the Pareto distribution decreases with increasing α

Figure 1 shows the relationship between α and the entropy of the Pareto distribution. 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. Intuitively, this would happen when a language consists of nearly only filler words and one meaningful word — if that word occurs exclusively, then it is not very informative in itself because there is no surprise, but if that word occurs rarely, then it is very informative but this contribution is lost in the average. However, as $\alpha \to 1$, the distribution becomes successively flatter, but maintaining a spike-like structure with a thick tail. Intuitively, this occurs when a small number of words are highly informative but all words have non-vanishing information content. As such, we expect that $\alpha = 1$ is near optimal when $c = \alpha - 1$ and that languages will have evolved (in the communicative sense, but following from the biological evolution) to have near optimal α.

4. Empirical Estimates Across Languages

Baixeries, Elvevg, and Ferrer-i-Cancho (2013) have previously shown that α decreases during first language acquisition on the basis of data from four Germanic languages, generally converging towards a value slightly below 1, with a fair

amount of inter-language variability. In the following, we examine α across a sample of 310 languages using the translations for Universal Declaration of Human Rights provided by the `nltk.corpus` Python package (Bird, Klein, & Loper, 2009) (see Table 1). We use ordinary least-squares regression to obtain estimates for the intercept ($\log c$) and slope ($-\alpha$) from Equation (2). Source code for the analysis can be found on Bitbucket.[c]

Table 1. Estimation of α by encoding. Error is standard error of the mean across single-language estimates. Encoding serves as a proxy for writing system; for this corpus, UTF-8 is typically used for ideographic scripts, while *Other* includes Hebrew and Arabic scripts.

Encoding	n	α
Latin1	190	0.90 ± 0.01
Cyrillic	10	0.74 ± 0.08
UTF8	86	0.97 ± 0.03
Other	110	0.93 ± 0.03
All	310	0.90 ± 0.01

4.1. *Constant of Proportionality*

In hypothesizing that $\alpha = 1$ is optimal, we assumed a proper Pareto distribution, i.e that $c = \alpha - 1$. Figure 2 shows that this is not quite true, with $\alpha - c = 0.8$ perhaps representing a more realistic assumption. As such, we expect that α will accordingly be shifted away from 1. In particular, we can consider accommodate this shift by setting α in Equation (4) equal to $\alpha' + k$ for some constant k. Then we have

$$P(x) = \frac{\alpha' + k - 1}{x^{\alpha' + k}}, \quad x \geq 1 \tag{8}$$

which implies that our estimate for α should be shifted away from one by the same amount as $\alpha - c$, i.e. we should expect $\alpha \approx 0.8$ to be near optimal.

4.2. *Exponent*

Figure 3 shows the distribution for estimates of α across languages, with a mean of about 0.9 (cf. Table 1). This is somewhat less than the original predicted idealized value of $\alpha = 1$; however, it is line with our updated estimate based on the bias in $\alpha - c$. Moreover, our c-corrected estimate provides an explanation of why previous work has found α to be near one, but rarely exactly one, even when corrected for observation error.

[c] `https://bitbucket.org/palday/evolang/`

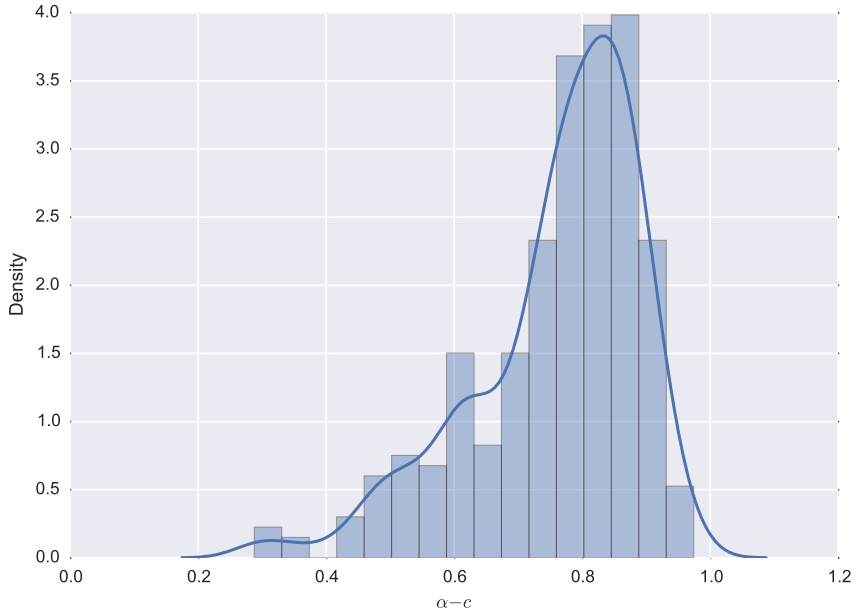

Figure 2. The difference between the constant of proportionality and α. If Zipfian distributions are exactly Pareto, then $\alpha - c = 1$, but this is not the case. Instead, the distribution is shifted left, with a mode of around 0.8

As writing systems may have an impact on blind orthographic measures (i.e. notions of "word" based purely on white-space delimited tokenization), we provide additional estimates divided by text encoding, which stands as a proxy for orthographic system, see Table 1 and Figure 4. Although the shape of the distribution varies across orthographic systems, the distributions all have a mode near 0.8, which suggests that the writing system does not lead to differences larger than those previously observed between closely related languages (cf. Baixeries et al., 2013).

5. Conclusion

Frequency-based explanations abound in empirical linguistics, from corpus linguistics to psycho- and neurolinguistics. Yet, they often suffer from a bit of a chicken and egg problem: X does this because X is more frequent, but how did X become more frequent in the first place? The results presented here provide a first step towards grounding empirical laws in the processing constraints and strategies of individual language users. We have shown how neurocomputational principles can motivate empirical laws via processing strategies, but not yet provided a direct

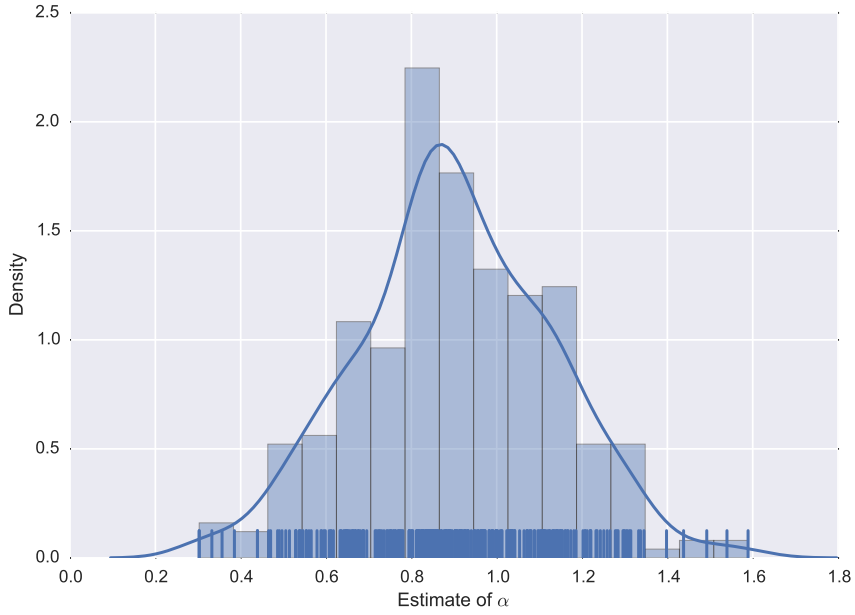

Figure 3. Distribution of α across languages. The surprisal-maximizing Pareto model predicts that $\alpha = 1$ should be ideal

derivation — our model is compatible with the principles but does not require neurobiological specifics and thus remains psychological. Nonetheless, we are able to formulate hypotheses in a principled way about the ideal values for parameters, which bear out in empirical testing. Having parameters that relate back to assumptions about basic cognitive strategies and processing constraints are far more valuable than parameters related to uninformed curve fitting. We can and should have both quantity and explanatory quality. Theories of language evolution need to be motivated by the biological entities doing the evolving.

Acknowledgements

I would like to thank Oliver Schallert for engaging discussions of empirical laws in synchrony and diachrony as well as two reviewers for invaluable suggestions on the textual presentation.

References

Baixeries, J., Elvevg, B., & Ferrer-i-Cancho, R. (2013). The evolution of the exponent of Zipf's law in language ontogeny. *PLoS One*, *8*(3), e53227.

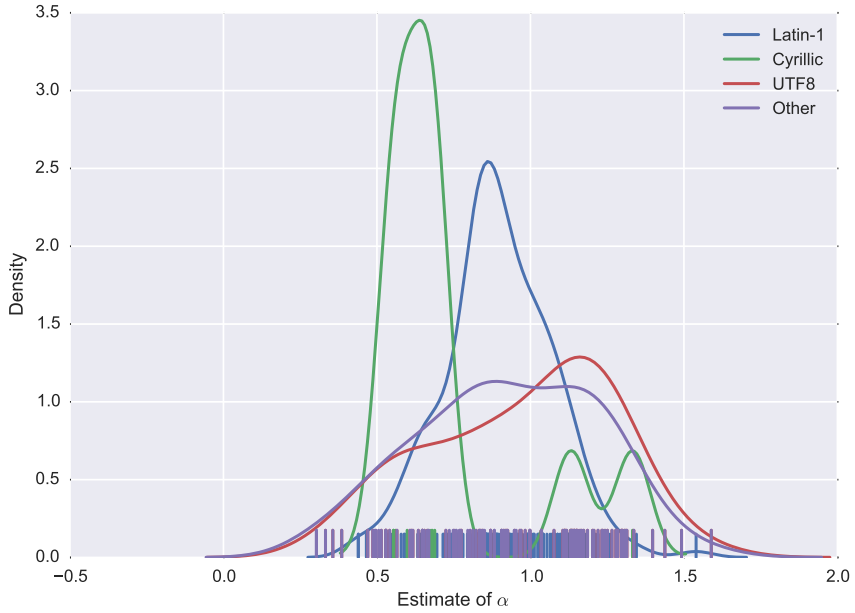

Figure 4. Distribution of α across languages. There is some variation in the shape of the distribution across languages, but the location seems similar with α near 0.8

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python.* O'Reily Media.

Ferrer-i-Cancho, R. (2015a). Reply to the commentary "be careful when assuming the obvious" by P.M. Alday. *Language Dynamics and Change*, *5*, 147–155.

Ferrer-i-Cancho, R. (2015b). The placement of the head that minimizes online memory: a complex systems approach. *Language Dynamics and Change*, *5*, 114–137.

Ferrer-i-Cancho, R., & Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, *100*(3), 788-791.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836.

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301.

Friston, K., Adams, R., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: saccades as experiments. *Frontiers in Psychology*, *3*(151).

Lieberman, E., Michel, J.-B., Jackson, J., Tang, T., & Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature*, *449*(7163),

713–716.
Pagel, M., Atkinson, Q. D., & Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout indo-european history. *Nature*, *449*(7163), 717–720.
Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*.
Zipf, G. K. (1929). Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology*, *40*, 1–95.
Zipf, G. K. (1935). *The psycho-biology of language.* Boston, MA: Houghton Mifflin Company.
Zipf, G. K. (1949). *Human behavior and the principle of least effort.* Cambridge, MA: Addison-Wesley.