A Zipfian Zetetic:
The Psycholinguistic Significance of the Zeta Function

Geoff Hollis & Chris Westbury
Department of Psychology, University of Alberta, Edmonton, AB, T6G 2E9 Canada.

WHAT IS A 'ZETETIC'?!

• 'Zetetic' is an obscure word, from a Greek root meaning 'to seek or enquire', that means "Investigation, inquiry (as in mathematics, etc.)". **In this poster we describe our investigation into the significance of the 1/f distribution in psycholinguistics.**

WHAT IS A '1/f DISTRIBUTION'?!

• A 1/f distribution is a distribution in which the frequency of a a thing is inversely proportional to its value. The most famous example, established by Zipf in 1949, is orthographic frequency. Highly frequent words are much less common than highly infrequent words, and the extent to which they are uncommon is extremely well described by the curve 1/frequency. This observation is now known as 'Zipf's law'. The 1/f distribution is often called a 'zeta distribution'.

INTRODUCTION

• The 1/f distribution is pervasive throughout the world. For instance, it describes the magnitude of earthquakes, city population sizes, and gold records attained by musical artists.

• A multitude of psycholinguistic variables also follow a 1/f distribution (Figure 1).

• Our zetetic addresses two main questions:
  • **Why is the 1/f (zeta) distribution so pervasive?**
  • **What implications might the zeta distribution have for our understanding of lexical organization?**

WHY IS A '1/f DISTRIBUTION' COMMON?

• We believe that **zeta distributions arise when an event's occurrence increases its probability of reoccurrence.**

• As a simple example, we demonstrate that given a set of choices where each choice's probability of being re-chosen after every selection increases, the frequencies of selection converge to follow a 1/x distribution (Figure 2).

• There appears to be some plausibility for this explanation in the structure and evolution of the lexicon. It is simple, and can tie together many disparate phenomena.

SHOULD WE CARE?

• We address the question of whether or not the Zeta distribution is of psycholinguistic interest by seeing how well distributional properties of Zeta-distributed lexical variables predict behavioral data.

Method

• Lexical decision reaction times (LDRTs) for 2390 words of length 4-6 were used from the English Lexicon Project (Balota, et al., 2002).

• Values for 6 Zeta-distributed lexical variables were calculated (Table 1).

• The R^2 values from correlations of two distributional properties (probability, cumulative probability) were compared to raw values, logged values, and a nonlinear transformation performed by our curve-fitting software, NUANCE (Hollis & Westbury, 2006; Hollis, Westbury & Peterson, in press).

Results

• In all but one case, the R^2 value of either probability or cumulative probability was approximately as good a predictor of LDRTs as the best predictor

• In two of the six cases, probability was the best predictor.

• These excellent fits raise the possibility that the lexicon is structured according using probability distributions

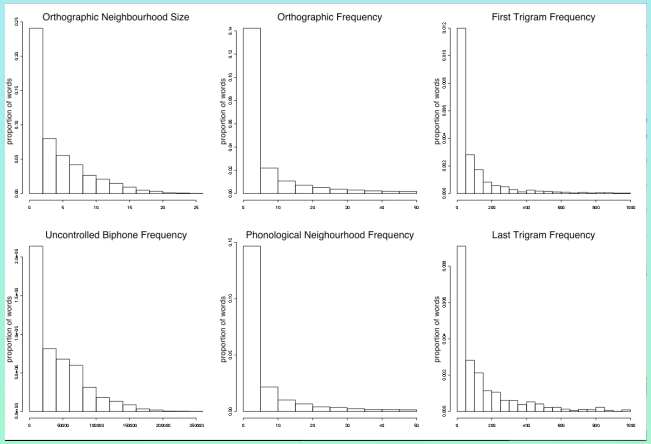

Figure 1. Many lexical variables follow Zeta distributions.

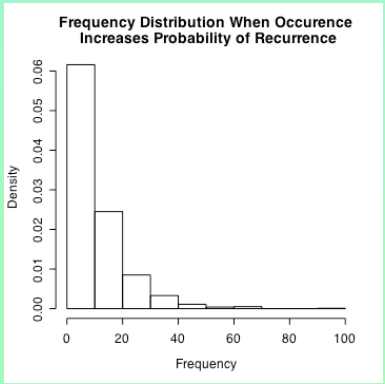

Figure 2. Data from a simple simulation demonstrating that when an event's selection increases its probability of re-selection, the resultant event frequencies follow a 1/x distribution.

Table 1.
Transformations for 6 zeta-distributed variables and their R^2 with LDRTs.

Predictor	Transformation					Best Predictor
	Linear	Logged	NUANCE	Prob.	Cum. Prob.	
OFREQ	0.02 x	0.33	0.36	0.33	0.36	NUANCE
ON	0.08	0.09	0.09	0.10	0.09	probability
FIRSTTRI	0.00 x	0.09	0.12	0.11	0.09 x	NUANCE
UNBP	0.01	0.01	0.02	0.00 x	0.01	NUANCE
PNFREQ	0.00 x	0.03 x	0.05	0.06	0.04	probability
LASTTRI	0.00 x	0.07 x	0.13	0.13	0.07 x	NUANCE

x indicates difference from highest R^2 value ($p < 0.05$)

Table 1: In all but one case, probability or cumulative probability of a factor is as good a transformation for predicting RTs as the best transformation we were able to find after extensive computational search .

CONCLUSION

• Processes that impinge on language processes often do so in nonlinear ways. One problem of accepting nonlinearity is that it expands the number of possible transformations of predictors to be evaluated; once the possibility of nonlinear transformation is presented, it is difficult to know what the best transformation is.
• We have shown here that many lexical variables show a 1/f distribution (Figure 1) and that probability and cumulative probability of these variables are often excellent predictors of RTs (Table 1).

• Log transformations have often been often used to maximize correlations of predictors with RTs: These are related because the cumulative probability of a zeta distribution is directly proportional to the log value (the integral of 1/x is ln(x)).
• There is reason to believe that probabilistic transformations may be 'better' transformations than the log transformations that estimate them:
i.) Our variables do not exactly follow Zeta distributions (there is some noise).
ii.) Log transformation are only as predictive or less predictive than our probabilistic transformations (Table 1)

• We still do not know why Zipf's law is so pervasive, but **we have presented evidence that Zeta distributions arise in systems that have feedback into themselves during development (Figure 2)** and we suggest that this is a plausible mechanism operating in the development and evolution of the lexicon.
• The human lexical system may be structured to take advantage of the 1/f distributional regularity, as suggested by the power of probability and cumulative probability to predict RTs

REFERENCES

Balota, D., Cortese, M., Hutchison, K., Neely, J., Nelson, D., Simpson, G., et al. (2002). *The english lexicon project: A web-based repository of descriptive and behavioral measures for 40,481 english words and nonwords*. http:// elexicon.wustl.edu.

Hollis, G., & Westbury, C. (2006). NUANCE: Naturalistic University of Alberta Nonlinear Correlation Explore. *Behavioral Research Methods, Instruments, and Computers*, 38, 8-23.

Hollis, G., Westbury, C., & Peterson, J. (in press). NUANCE 3.0: Using Genetic Programming to Model Variable Relationships. *Behavioral Research Methods, Instruments, and Computers*.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley, Reading, MA, 1949.

Acknowledgements: This work was supported by the Natural Sciences and Engineering Research Council of Canada.