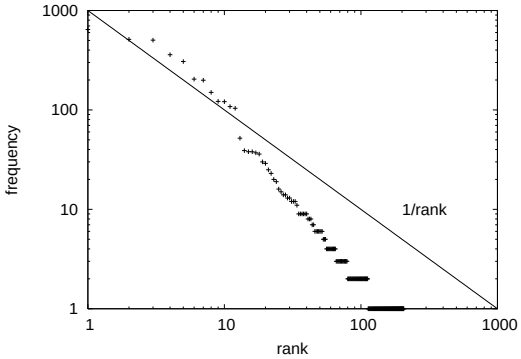
Generalisation as a bias toward the emergence of Zipf's law

Paul Vogt, Tilburg University, The Netherlands, p.a.vogt@uvt.nl

All human languages show the characteristics of Zipf's law (Zipf, 1949). This law is the observation that the frequency with which words occur decays as a power law of their rank. If words are decreasingly ordered based on their frequency, the frequency $f(k)$ of a word with rank k is given by $f(k) \propto k^{-\alpha}$, where $\alpha \approx 1$. Zipf explains the emergence of his law using the principle of least effort; speakers want to minimise the effort for producing utterances, and hearers want to minimise the effort of understanding them. This principle seems to hold at various levels, such as the phonological and the lexical level (Ferrer i Cancho and Solé, 2003).

In this paper, I show how Zipf's law can emerge through a tendency to minimise the effort in categorising perceptual features. I do not claim that this is the only bias at work; other biases, such as influences taken from the environment and discourse models, have shown a tendency toward Zipfian distributions as well (Tullo and Hurford, 2003). However, the principle of least effort – on which the findings in this paper are based – appears to be sound (Ferrer i Cancho and Solé, 2003).

The data presented in this paper are drawn from robotic experiments that have been carried out at the end of the past century, e.g., (Vogt, 2000). In these experiments, two mobile robots developed a shared lexicon from scratch of which the words' meanings were grounded by the robots' interactions with the environment. The experiments were based on adaptive language games (Steels, 1996) in which a speaker produces an utterance, which the hearer tries to interpret. During the experiments, the robots developed categories (meanings) in a number of conceptual spaces that had various levels of granularity, but were spanned by the same quality dimensions (perceptual feature dimensions). Thus, sparsely filled conceptual spaces contained more general categories than densely filled ones. In the language games, the robots first tried to categorise the perceptual features in the sparsely filled spaces. When they failed, they incrementally tried the more specialised spaces. This way, the robots preferred to communicate about more general concepts than less general ones. The main reason for this design was to minimise computational complexity (finding categories in densely filled spaces is computationally expensive).



freq.	≥ 50	≥ 10	≥ 5	≥ 2	$= 1$
%	100	67	38	26	19
layer	1.00	1.86	2.57	3.21	2.87

Recent (re)inspection of the data revealed the emergence of a Zipfian distribution in the relation between word-frequencies and their rank (see figure). Closer inspection even showed that word-meanings about general categories appeared more frequently than those about specialised categories (see % row in table, which gives the percentages of word-meanings of which the meanings are at the most general layer). Categories emerged at five different layers of varying granularity, the final row of the table shows the average layer with respect to the frequencies of word-meanings (layer 1 being the most general).

Given these results, I conclude that having a tendency to minimise the effort by trying to communicate using the most general meaning in a situation, has led for these robotic experiments to the emergence of a Zipfian distribution in word-meaning rankings. Hence, I hypothesise that a generalisation bias – as a strategy imposed by the principle of least effort – leads, among other biases, to a Zipfian distribution in natural languages.

References

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

L. Steels. 1996. Emergent adaptive lexicons. In P. Maes, editor, *From Animals to Animats 4: Proceedings of the Fourth International Conference On Simulating Adaptive Behavior*, Cambridge Ma. The MIT Press.

C. Tullo and J. R. Hurford. 2003. Modelling Zipfian distributions in language. In S. Kirby, editor, *Language Evolution and Computation, Proceedings of the workshop at ESSLLI.*

P. Vogt. 2000. Bootstrapping grounded symbols by minimal autonomous robots. *Evolution of Communication*, 4(1):89–118.

G. K. Zipf. 1949. *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology.* Addison-Wesley, Cambridge, MA.