Phonotactics is affected by statistical scaling laws less than the lexicon

Andreas Baumann, Theresa Matzinger, Kamil Kaźmierski
University of Vienna, Adam Mickiewicz University Poznań
andreas.baumann@univie.ac.at, theresa.matzinger@univie.ac.at, kamil.kazmierski@wa.amu.edu.pl

Keywords: complex systems, Heaps' law, Zipf's law, phonotactics, diversity, inventory size

Lexical systems have been shown to follow statistical laws characteristic of many complex systems (Corominas-Murtra & Solé 2010; Ferrer-i-Cancho 2016). Most prominently, Zipf's law models the inverse relationship between word frequency and rank (Zipf 1949). The law is hypothesized to be related to several other statistical patterns that one can observe in language such as the inverse relationship between word frequency and word length (Baayen, 2001). Another prominent law potentially linked to Zipf's law is Heaps' law (Heaps, 1978). It is a model of a system's complexity (e.g. the number of word types) depending on the number of tokens in it, i.e. corpus size. According to Heaps' law, complexity grows sublinearly with the number of tokens in a corpus. These statistical scaling laws have been hypothesized to be a consequence of many factors, among others cognitive determinants such as limited memory, communicative efficiency, and semantic organization (Piantadosi 2014).

Zipf's law applies less strongly to phonology. For many languages it has been shown that the relationship between phoneme frequency and phoneme rank is only roughly modeled by Zipf's law (Tambovtsev & Martindale 2007), although it still applies to the relationship between phoneme duration and frequency (Kuperman, Ernestus & Baayen 2008). In this paper, we focus on the domain of phonotactics, i.e. sequences of sounds, which is – put into simplified terms – located between phonology and the lexicon. Phonotactics covers longer items than phonology, but phonotactic items (or n-phones) carry much less meaning than lexical items do (although phonotactic items can have sound-symbolic properties and/or fulfil functional roles). Research done on statistical laws in phonotactics is relatively limited (but see e.g. Ha, Hanna, Ming, & Smith, 2009; Kuperman et al., 2008). We provide a systematic analysis of Zipf's law and Heaps' law in phonotactic systems. Based on a corpus of spoken English (Buckeye), we estimate scaling-law exponents for phonotactic items of different length and in two different conditions (within-word and within-and-across word phonotactics). We measure phonotactic complexity both in terms of inventory size and with frequency-based diveristy measures (Hill 1973).

We find that phonotactics is less strongly affected by the inspected scaling laws than this is the case for the lexicon. Furthermore, we show that phonotactic length has a crucial impact on Heaps' law in phonotactics. For phonotactic sequences of length 6 (which roughly equals the average phonological length of words), Heaps' exponent is about 0.8 (which approximates lexical estimates of Heaps' exponent). In contrast, Zipf's law is not significantly affected by the length of phonotactic items. We suggest that cognitive constraints apply less to the phonotactic domain, partially because phonotactic items carry much less meaning, which in turn affects cognitive constraints related to memory and semantic organization (Piantadosi 2014).

References

Baayen, R Harald. 2001. Word frequency distributions. Dordrecht: Kluwer Academic Publishers.
Corominas-Murtra, Bernat & Ricard V. Solé. 2010. Universality of Zipf's law. Physical Review E..
Ferrer-i-Cancho, Ramon. 2016. Compression and the origins of Zipf's law for word frequencies. Complexity. doi:10.1002/cplx.21820.
Ha, Le Quan, Philip Hanna, Ji Ming & F. J. Smith. 2009. Extending Zipf's law to n-grams for large corpora. Artificial Intelligence Review. doi:10.1007/s10462-009-9135-4.
Heaps, H. S. 1978. Information retrieval: computational and theoretical aspects. New York, San Francisco, London, Academic Press, 1978, 344p.s.
Hill, Mark. 1973. Diversity and evenness. Ecology 54(2). 427–432.
Kuperman, Victor, Mirjam Ernestus & Harald Baayen. 2008. Frequency distributions of uniphones, diphones, and triphones in spontaneous speech. The Journal of the Acoustical Society of America 124. 3897–3908.
Piantadosi, Steven T. 2014. Zipf's word frequency law in natural language: A critical review and future directions. Psychonomic Bulletin and Review. doi:10.3758/s13423-014-0585-6.
Tambovtsev, Yuri & Colin Martindale. 2007. Phoneme Frequencies Follow a Yule Distribution. SKASE Journal of Theoretical Linguistics [online].
Zipf, Georg Kingsley. 1949. Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology. Cambridge, MA.: Addison-Wesley Press.