THE EVOLUTION OF OPTIMIZED LANGUAGE IN THE LIGHT OF STANDARD INFORMATION THEORY

Ramon Ferrer-i-Cancho*[1] and Christian Bentz[2,3]

*Corresponding author, rferrericancho@cs.upc.edu
[1]Department of Computer Science, Universitat Politècnica de Catalunya, Barcelona, Catalonia
[2]Department of General Linguistics, University of Tübingen, Tübingen, Germany
[3]DFG Center for Advanced Studies, University of Tübingen, Tübingen, Germany

Zipf's law of abbreviation, the tendency of more frequent words to be shorter, emerges as a universal property of languages (Bentz & Ferrer-i-Cancho, 2016). Language users have been shown to adhere to it as a result of combining two conditions: accuracy, i.e. avoiding ambiguity, and "efficiency", i.e. using word forms as short as possible (Kanwal, Smith, Culbertson, & Kirby, 2017). Random typing has been suggested as a non-functional alternative to statistical laws of language (Miller & Chomsky, 1963; Li, 1992). However, a caveat of random typing is that it is not purely non-functional from an information theoretic perspective: random typing can be seen as a case of optimal nonsingular encoding of information (Ferrer-i-Cancho, Bentz, & Seguin, 2015). This is an example of how information theory and evolutionary linguistics can develop mutually enriching connections.

Another example is the striking similarity between the two conditions above and coding theory, where the problem of compression is the problem of minimizing L, the mean length of codes (e.g., words) under some coding scheme (Cover & Thomas, 2006). The minimization of L matches the "efficiency" condition, whereas nonsingular coding is a coding scheme that matches beautifully the accuracy condition. Nonsingular coding is equivalent to using unambiguous words to represent meanings (Cover & Thomas, 2006, p. 105). Extensions of standard information theory predict that in case of optimal coding (maximum "efficiency" and maximum accuracy), the correlation between word frequency and word length *cannot* be positive and, in general, it is expected to be negative in concordance with Zipf's law of abbreviation (Ferrer-i-Cancho et al., 2015).

Given the alphabet of a language and the probabilities of the word types, we can calculate L_{min}, the minimum mean word length that can be achieved assuming a certain scheme (Ferrer-i-Cancho et al., 2015). Then, we can measure the degree of optimality of a language with $\eta = L_{min}/L$. η is the so-called coding efficiency (Borda, 2011), and ranges between 0 and 1, reaching 1 in case of an optimal communication system.

DOI:10.12775/3991-1.029

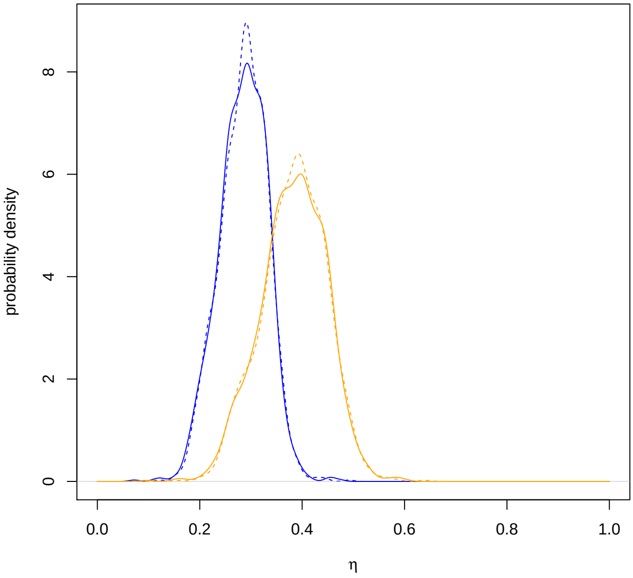

Figure 1. The distribution of η in the PBC under nonsingular coding (blue) and uniquely decodable coding (orange). Dashed lines correspond to a prefix of 10^5 tokens of the original text to reduce the estimation bias due to differences in text length.

Preliminary analyses of more than 1000 languages in the PBC, the Parallel Bible Corpus (Mayer & Cysouw, 2014), suggest that real languages are optimized to a 30% average (Fig. 1). Interestingly, the average optimization ratio increases to 40% (Fig. 1) if the nonsingular coding scheme is replaced by uniquely decodable encoding, which requires not only that there is no ambiguity between word types as in nonsingular coding, but also that a concatenation of letters without blanks allows for just one segmentation into a sequence of word tokens.

Such a mixture of suboptimalities in languages, which are neither perfectly nonsingular nor perfectly uniquely decodable, provides support for the hypothesis that Zipf's law for word frequencies stems from a competition between optimal nonsingular coding and optimal uniquely decodable coding (Ferrer-i-Cancho, 2016). This account is not just one more model of Zipf's law. Compression also predicts Zipf's law of abbreviation as reviewed above, as well as Menzerath's law (Gustison, Semple, Ferrer-i-Cancho, & Bergman, 2016). Hence, it illustrates the combination of predictive power, parsimony and mathematical rigor that information theory offers to understand how languages evolve universal properties.

Acknowledgements

R.F.C is funded by the grants 2014SGR 890 (MACDA) from AGAUR (Generalitat de Catalunya) and the grant TIN2017-89244-R from MINECO. C. B. is funded by the German Research Foundation (DFG FOR 2237) and by the ERC Advanced Grant 324246 EVOLAEMP.

References

Bentz, C., & Ferrer-i-Cancho, R. (2016). Zipf's law of abbreviation as a language universal. In C. Bentz, G. Jger, & I. Yanovich (Eds.), *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics.* University of Tbingen.

Borda, M. (2011). *Fundamentals in information theory and coding.* Berlin: Springer.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory.* New York: Wiley. (2nd edition)

Ferrer-i-Cancho, R. (2016). Compression and the origins of Zipf's law for word frequencies. *Complexity*, *21*, 409-411.

Ferrer-i-Cancho, R., Bentz, C., & Seguin, C. (2015). Compression and the origins of Zipf's law of abbreviation. *http://arxiv.org/abs/1504.04884*.

Gustison, M. L., Semple, S., Ferrer-i-Cancho, R., & Bergman, T. (2016). Gelada vocal sequences follow Menzerath's linguistic law. *Proceedings of the National Academy of Sciences USA*, *13*, E2750-E2758.

Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017). Zipfs law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, *165*, 45-52.

Li, W. (1992). Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE T. Inform. Theory*, *38*(6), 1842-1845.

Mayer, T., & Cysouw, M. (2014). Creating a massively parallel bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014* (pp. 3158–3163).

Miller, G. A., & Chomsky, N. (1963). Finitary models of language users. In R. D. Luce, R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 2, p. 419-491). New York: Wiley.