Human languages trade off complexity against efficiency

Alexander Koplenig[1], Sascha Wolfer[1] & Peter Meyer[1]
[1] Department of Lexical Studies, Leibniz Institute for the German Language (IDS), Mannheim, Germany, koplenig@ids-mannheim.de

Keywords: language complexity, language efficiency, information theory, quantitative typology, quantitative linguistics

A central goal of linguistics is to understand the diverse ways in which human language can be organized (Gibson et al. 2019; Lupyan & Dale 2016). In our contribution, we present results of a large scale cross-linguistic analysis of the statistical structure of written language. To this end, we have trained a language model on more than 6,500 different documents as represented in 41 parallel/multilingual corpora consisting of ~3.5 billion words or ~9.0 billion characters and covering 2,069 different languages that are spoken as a native language by more than 90% of the world population or ~46% of all languages that have a standardized written representation. Figure 1 shows that our database covers a large variety of different text types, e.g. religious texts, legalese texts, subtitles for various movies and talks, newspaper texts, web crawls, Wikipedia articles, or translated example sentences from a free collaborative online database. Furthermore, we use word frequency information from the Crúbadán project that aims at creating text corpora for a large number of (especially under-resourced) languages (Scannell 2007). We statistically infer the entropy rate of each language model as an information-theoretic index of (un)predictability/complexity (Schürmann & Grassberger 1996; Takahira, Tanaka-Ishii & Dębowski 2016). Equipped with this database and information-theoretic estimation framework, we first evaluate the so-called 'equi-complexity hypothesis', the idea that all languages are equally complex (Sampson 2009). We compare complexity rankings across corpora and show that a language that tends to be more complex than another language in one corpus also tends to be more complex in another corpus. This constitutes evidence against the equi-complexity hypothesis from an information-theoretic perspective. We then present, discuss and evaluate evidence for a complexity-efficiency trade-off that unexpectedly emerged when we analysed our database: high-entropy languages tend to need fewer symbols to encode messages and vice versa. Given that, from an information theoretic point of view, the message length quantifies efficiency – the shorter the encoded message the higher the efficiency (Gibson et al. 2019) – this indicates that human languages trade off efficiency against complexity. More explicitly, a higher average amount of choice/uncertainty per produced/received symbol is compensated by a shorter average message length. Finally, we present results that could point toward the idea that the absolute amount of information in parallel texts is invariant across different languages.

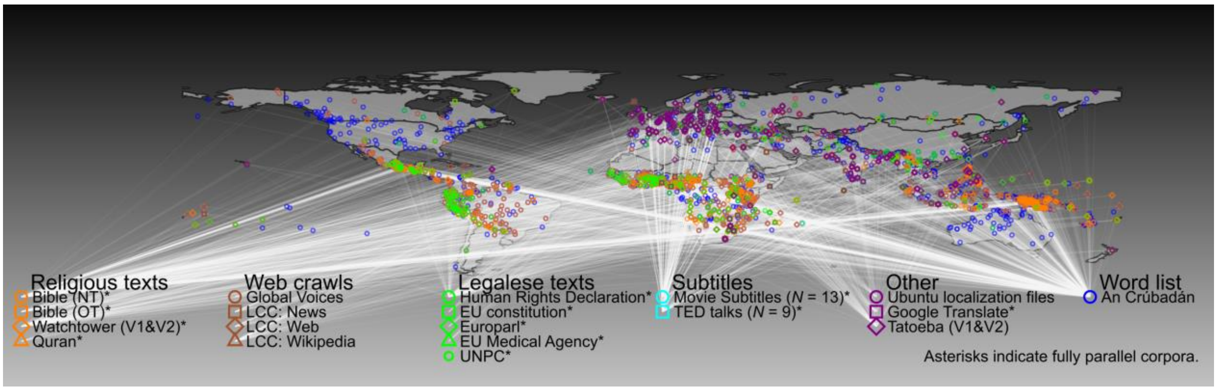

Fig. 1: Collected corpora and their geographical distribution.

References

Gibson, Edward, Richard Futrell, Steven T. Piandadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen & Roger Levy. 2019. How Efficiency Shapes Human Language. *TRENDS in Cognitive Science* 23(5). 389–407. https://doi.org/10.1016/j.tics.2019.02.003.

Lupyan, Gary & Rick Dale. 2016. Why Are There Different Languages? The Role of Adaptation in Linguistic Diversity. *TRENDS in Cognitive Science* 20(9). 649–660. https://doi.org/10.1016/j.tics.2016.07.005.

Sampson, Geoffrey. 2009. A linguistic axiom challenged. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language complexity as an evolving variable*, 1–18. Oxford: Oxford University Press.

Scannell, Kevin P. 2007. The Crúbadán Project: Corpus building for under-resourced languages. In *Proceedings of the 3rd Web as Corpus Workshop: Building and Exploring Web Corpora*, vol. 4, 5–15. http://cs.slu.edu/~scannell/pub/wac3.pdf.

Schürmann, Thomas & Peter Grassberger. 1996. Entropy estimation of symbol sequences. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 6(3). 414. https://doi.org/10.1063/1.166191.

Takahira, Ryosuke, Kumiko Tanaka-Ishii & Łukasz Dębowski. 2016. Entropy Rate Estimates for Natural Language—A New Extrapolation of Compressed Large-Scale Corpora. *Entropy* 18(10). 364. https://doi.org/10.3390/e18100364.