Originallypublishedin:Trawinski,Beata/Kupietz, Marc/Proost,Kristel/Zinken,Jörg(eds.): 10thInternationalContrastiveLinguisticsConference(ICLC-10),18-21July,2023, Mannheim,Germany.BookofAbstracts.– Mannheim: IDS-Verlag,2023.S.190-192
DOI:https://doi.org/10.14618/f8rt-m155

Alexander Koplenig/Sascha Wolfer/Peter Meyer

HUMAN LANGUAGES TRADE OFF COMPLEXITY AGAINST EFFICIENCY

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 (Koplenig/Wolfer/Meyer 2023) we approach this question from an information-theoretic perspective. To this end, we conduct a large scale quantitative cross-linguistic analysis of written language by training a language model on more than 6,500 different documents as represented in 41 multilingual text collections, so-called 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. We statistically infer the entropy of each language model as an index of (un. 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 indexof (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.

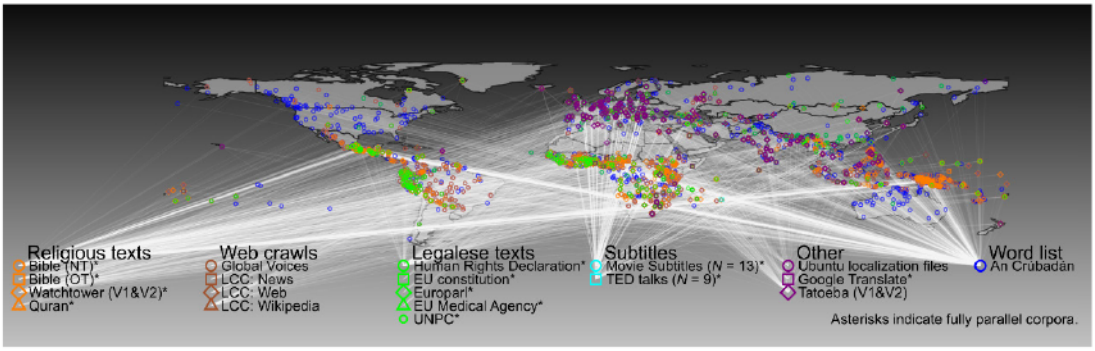

Fig. 1: Collected corpora and their geographical distribution

References

Gibson, Edward/Futrell, Richard/Piandadosi, Steven T./Dautriche, Isabelle/Mahowald, Kyle/Bergen, Leon/Levy, Roger (2019): How efficiency shapes human language. In: TRENDS in Cognitive Science 23 (5), pp. 389–407. https://doi.org/10.1016/j.tics.2019.02.003.

Koplenig, Alexander/Wolfer, Sascha/Meyer, Peter (2023): Human languages trade off complexity against efficiency. Preprint. In: Research Square. https://doi.org/10.21203/rs.3.rs-1462001/v2.

Lupyan, Gary/Dale, Rick (2016): Why are there different languages? The role of adaptation in linguistic diversity. In: TRENDS in Cognitive Science 20 (9), pp. 649–660. https://doi.org/10.1016/j.tics.2016.07.005.

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

Scannell, Kevin P. (2007): The Crúbadán Project: Corpus building for under-resourced languages. In: Fairon, Cédric/Naets, Hubert/Kilgarriff, Adam/de Schryver, Gilles-Maurice (eds.): Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop. Louvain: Presses universitaires de Louvain, pp. 5–15. http://cs.slu.edu/~scannell/pub/wac3.pdf.

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

Takahira, Ryosuke/Tanaka-Ishii, Kumiko/Dębowski, Łukasz (2016): Entropy rate estimates for natural language – a new extrapolation of compressed large-scale corpora. In: Entropy 18 (10), p. 364. https://doi.org/10.3390/e18100364.

Contact information

Alexander Koplenig
Department of Lexical Studies, Leibniz Institute for the German Language (IDS), Mannheim, Germany
koplenig@ids-mannheim.de

Sascha Wolfer
Department of Lexical Studies, Leibniz Institute for the German Language (IDS), Mannheim, Germany
wolfer@ids-mannheim.de

Peter Meyer
Department of Lexical Studies, Leibniz Institute for the German Language (IDS), Mannheim, Germany
meyer@ids-mannheim.de

Bibliographical information

This text is part of the publication: Trawiński, Beata/Kupietz, Marc/Proost, Kristel/Zinken, Jörg (eds.) (2023): 10. International Contrastive Linguistics Conference (ICLC). Book of Abstracts (pre-conference version). Mannheim: IDS-Verlag.