Beyond words: Lower and upper bounds on the entropy of subword units in diverse languages

Christian Bentz[1]
[1]University of Tübingen, chris@christianbentz.de

Keywords: Information theory, language complexity, word meanings

Languages are diverse. At any level of structure, from phonemes to discourse, we find a multitude of encoding strategies used across the circa 7000 thousand spoken and signed languages of the world. A crucial step to understand this diversity is to map out the space of possible languages. One dimension of this space is the diversity of words used to encode different meanings (cf. Gibson et al., 2017; Majid et al., 2018; Zaslavsky et al., 2018).

From an information-theoretic perspective, the encoding potential of words – their entropy (H) – constitutes an upper-bound on the mutual information (MI) with their meanings. Assume a set of words W and a set of meanings M. It follows from standard information theory that (Ferrer-i-Cancho & Díaz-Guilera, 2007, p. 13)

$$MI(W,M) \leq H(W). \tag{1}$$

In other words, the entropy is a fundamental restriction on (unambiguous) information transfer. While for natural languages it is hard to realistically estimate the mutual information between words and meanings, the entropy $H(W)$ can be estimated based on written material and transliterated spoken language.

Orthographic word entropies have been recently estimated for diverse languages and texts (Montemurro & Zanette, 2011; Bentz et al., 2017; Koplenig et al., 2017). Non-trivial lower and upper bounds on word entropies emerge from this research. These bounds are likely related to the trade-off between ease of learning and expressiveness. It has recently been shown that low word entropy distributions facilitate language learning in children (Lavi-Rotbain & Arnon, 2022). High word entropy, on the other hand, facilitates rapid information transfer (Ferrer-i-Cancho & Díaz-Guilera, 2007). At the level of words, natural languages are neither perfectly learnable, nor perfectly expressive. Their encoding potential is the outcome of both pressures acting simultaneously.

However, there are several recurring question with regards to this information-theoretic research:

- Since orthographic words are (somewhat) arbitrary units of writing (Haspelmath, 2011; Wray, 2015), what happens to these bounds if we steer away from orthographic words, and rather use subword patterns as units?
- Are written and spoken languages located in different areas of the entropy space?
- Does higher entropy on the signal side of a language (i.e. words or subwords) imply higher entropy of this language in general?

In this talk, these questions are addressed based on current research.

References

Bentz, Christian, Dimitrios Alikaniotis, Michael Cysouw & Ramon Ferrer-i Cancho. 2017. The entropy of words—learnability and expressivity across more than 1000 languages. *Entropy* 19(6). 275.

Ferrer-i-Cancho, Ramon & Albert Díaz-Guilera. 2007. The global minima of the communicative energy of natural communication systems. *Journal of Statistical Mechanics: Theory and Experiment* 2007(06). P06009.

Gibson, Edward, Richard Futrell, Julian Jara-Ettinger, Kyle Mahowald, Leon Bergen, Sivalogeswaran Ratnasingam, Mitchell Gibson, Steven T Piantadosi & Bevil R Conway. 2017. Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences* 114(40). 10785–10790.

Haspelmath, Martin. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia linguistica* 51(s1000). 31–80.

Koplenig, Alexander, Peter Meyer, Sascha Wolfer & Carolin Müller-Spitzer. 2017. The statistical trade-off between word order and word structure–large-scale evidence for the principle of least effort. *PloS one* 12(3). e0173614.

Lavi-Rotbain, Ori & Inbal Arnon. 2022. The learnability consequences of zipfian distributions in language. *Cognition* 223. 105038.

Majid, Asifa, Seán G Roberts, Ludy Cilissen, Karen Emmorey, Brenda Nicodemus, Lucinda O'grady, Bencie Woll, Barbara LeLan, Hilário De Sousa, Brian L Cansler et al. 2018. Differential coding of perception in the world's languages. *Proceedings of the National Academy of Sciences* 115(45). 11369–11376.

Montemurro, Marcelo A & Damián H Zanette. 2011. Universal entropy of word ordering across linguistic families. *PLoS One* 6(5). e19875.

Wray, Alison. 2015. Why are we so sure we know what a word is?, 725–750. Oxford University Press.

Zaslavsky, Noga, Charles Kemp, Terry Regier & Naftali Tishby. 2018. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences* 115(31). 7937–7942.