Conditional Word Count and Huffman Code Size are Two Sides of the Same Coin: Response to Koplenig

Pedro Aceves[1]
James A. Evans[2,3]

[1] Department of Management and Organization, Johns Hopkins University, Baltimore, MD 21202, paceves@jhu.edu
[2] Department of Sociology & Knowledge Lab, University of Chicago, 5735 South Ellis Avenue, Chicago, IL 60637. jevans@uchicago.edu
[3] Santa Fe Institute, 1399 Hyde Park Rd, Santa Fe, NM 87501

In a response to Aceves and Evans[1], Koplenig[2] argues that once the influence of corpus size (by which he means the length of the document in number of words) is accounted for, the relationship between information density and semantic density not only disappears but slightly reverses. His argument relies on conceptualizing corpus size and information density as distinct theoretical constructs. In fact, they are one and the same, and both capture the density of information in language.

The conceptual foundation of Aceves and Evans[1] is that information density as a theoretical construct is concerned with how information is distributed across the units of a language, which in their case is words. One can imagine several ways to represent this theoretical construct empirically. Our proposed approach was laid out in our paper and uses the Huffman encoding of the word distribution within a document. An alternative approach, which is less principled (as it does not take the distribution of words into account), is to simply count the number of words within a document, conditional on the same content. Because the information content of the parallel translations is the same across comparison languages, we can infer that the more words present within a document covering the same material, the less information is encoded in each. Koplenig's analysis of New Testament parallel translations in isolation demonstrates this clearly.

Thus, word count information density (what Koplenig calls corpus size) and Huffman information density are mirror images of one another. One might expect that they would be highly correlated with each other because they represent the same theoretical construct. This is in fact the case. When word count is standardized for each parallel translation (i.e., each English-other_language pair)[1], the correlation between Huffman encoded information density and word count information density is 0.94 ($p < 0.001$, Fig. 1c).

[1] We do this because even within the same corpus, information content varies, often considerably, depending on the language that is being compared to its English equivalent. We create this ratio measure the same way that we did for the Huffman size measure of information density, with English being the denominator for each parallel translation and the pair language being the numerator.

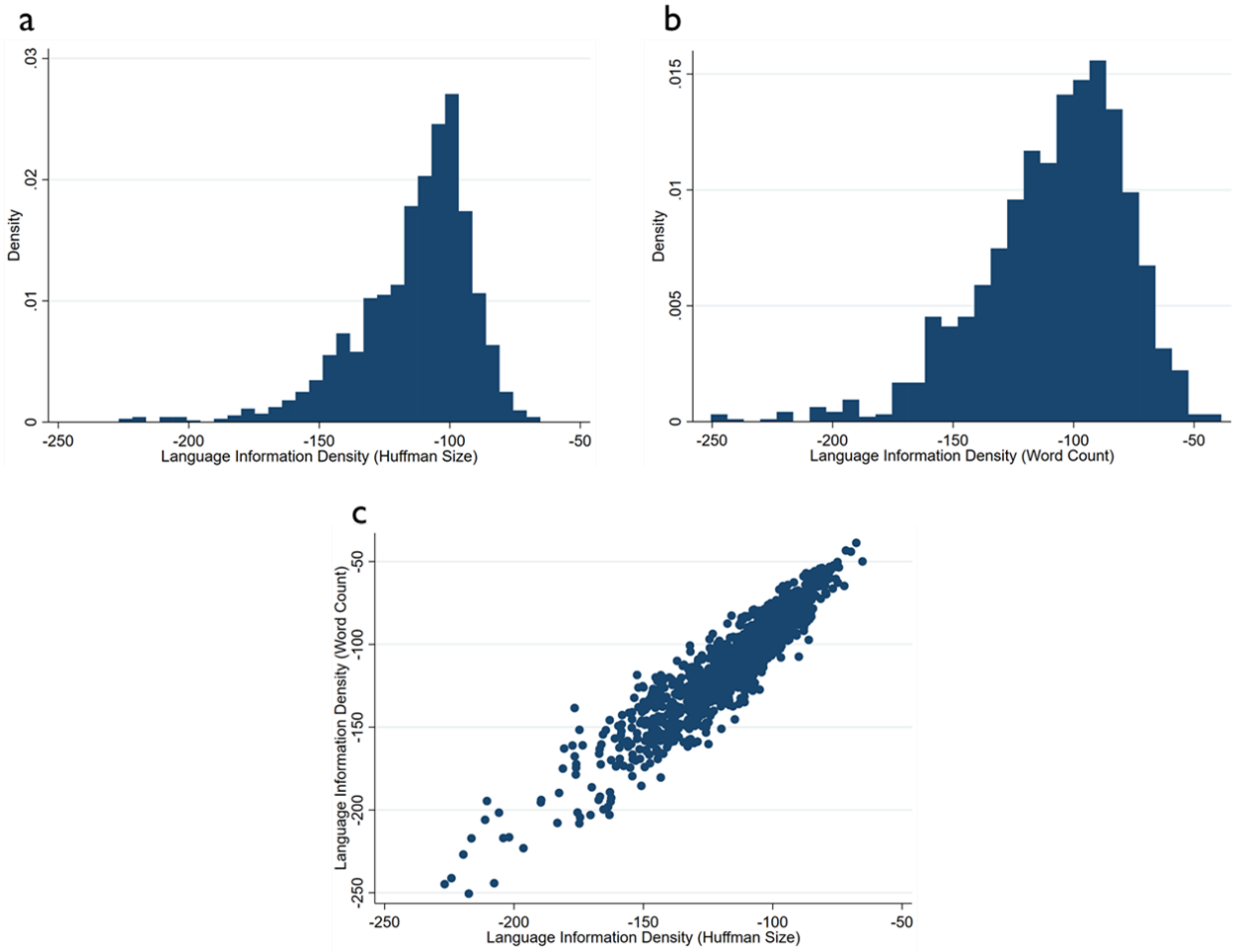


Fig. 1 | Distribution of Huffman size information density and word count information density, and their correlation. a, Distribution of Huffman size information density. **b,** Distribution of word count information density. **c,** scatterplot of Huffman size information density and word count information density.

Because these two variables represent the same theoretical construct, it is reasonable to expect the same relationship, regardless of which variable is used in the analysis. Figure 2 shows this is the case. A one-unit increase in Huffman size information density is associated with a 1.00-unit increase in conceptual space density (95% confidence interval (CI), 0.82, 1.18; $P < 0.001$; Fig. 2a). A one-unit increase in word count information density is associated with a 0.87-unit increase in conceptual space density (95% confidence interval (CI), 0.75, .98; $P < 0.001$; Fig. 2b).

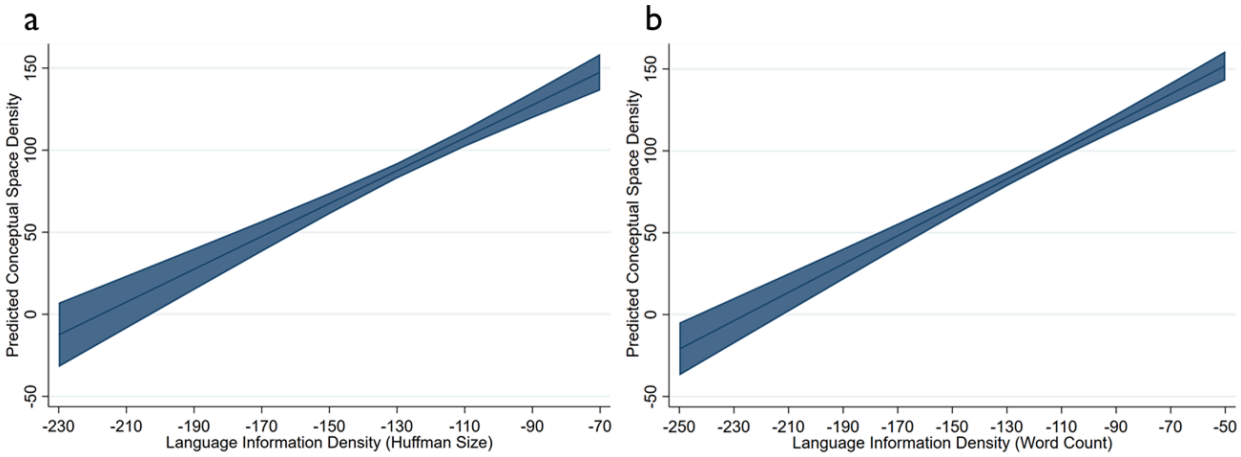


Fig. 2 | Relation between each information density variable and semantic density. a, Association between Huffman size information density and semantic density. 95% confidence intervals represented by the shaded region. Semantic density is the outcome of a three-level mixed model with random intercepts at both the language-family and language-within-language-family levels. Controls for corpus and corpus size are included. **b,** Association between word count information density and semantic density. 95% confidence intervals represented by the shaded region. Semantic density is the outcome of a three-level mixed model with random intercepts at both the language-family and language-within-language-family levels. Controls for corpus and corpus size are included.

The relationship between Huffman coding and number of words, conditional on the same semantic content is further recommended by their theoretical association with Shannon information entropy *H*, which indicates a sequence of events with less predictability, more randomness, and less compressibility. Shannon entropy

$$H = \sum_{i=1}^{n} mp_i \log \frac{1}{p_i}$$

where n represents possible outcomes, with probabilities $p_1, p_2, \ldots, p_n$, and m represents the length of the evaluated sequence. In our context, m could be the number of words (or characters) in the corpus, such that more words (or a longer alphabet), conditional on the same content (e.g., New Testament, movie subtitles), will yield a higher H. Shannon entropy defines the lower bound of compressibility such that no encoding—Huffman included—can compress the sequence in fewer bits. Nevertheless, we used Huffman encoding because it takes advantage of event (e.g., word) co-occurrences and assigns bits accordingly. In sum, conditional on content, more words will yield higher entropy, which will indicate lower information density. Our findings confirm that number of words, entropy, and Huffman code length—all manifestations of information density (e.g., more information is passing through a fixed number of words)—demonstrate strong and positive relationships with semantic density.

In summary, Huffman coding and number of words, conditional on content, measure the same underlying construct, which is how linguistic symbols are differentially used to encode information content. Our analysis hinges on an information fixed-effect strategy. Every

comparison we make holds the content or information constant across the parallel translations (i.e., the same religious, government, technical, economic, and cultural information is compared between English and its parallel translation pair).[2] Results provided by Koplenig take the information density out of information density. In doing so, they miss our findings that across its several measurements, information density is associated with (1) faster communication of the same content, but paradoxically (2) less information within conversations and articles when content is not held constant. Informationally dense languages cycle through the same content in more ways, and with potentially greater collective precision.

References

1. Aceves, P. & Evans, J. A. Human languages with greater information density have higher communication speed but lower conversation breadth. *Nat Hum Behav* (2024) doi:10.1038/s41562-024-01815-w.
2. Koplenig, A. Corpus size strongly matters when analysing word frequency distributions. doi:10.31219/osf.io/p5nhd.

[2] In private correspondence, Koplenig clarified his misunderstanding of our fixed effect strategy by stating that "for the Subs16 corpus,...there are some languages, e.g., Esperanto, with available information for less than 100 movie subtitles, while there are thousands of available subtitles for other languages, e.g. French, German, …" Our fixed effect approach would *only* compare Esperanto and English on subtitles about *the same* movies and shows. It would obviously compare German and English on far more. We note that we included the raw number of words as a control variable in order to compare across large and small corpora with different information (e.g., the New Testament, Subtitles, and United Nations proceedings). We see that this may have been mistakenly confused with controlling for the number of words, conditional on content. In our paper, we directly accounted for the Huffman coding score for each language, relative to a comparable code *of the same content* in English. As we show above, when we include relative number of words rather than relative Huffman code size, we obtain the same effect.