Spotlight

Information density as a predictor of communication dynamics

Gary Lupyan [ID], [1,*]
Pablo Contreras Kallens [ID], [2,*] and
Rick Dale [ID] [3,*]

In a recent paper, Aceves and Evans computed information and semantic density measures for hundreds of languages, and showed that these measures predict the pace and breadth of ideas in communication. Here, we summarize their key findings and situate them in a broader debate about the adaptive nature of language.

The thousands of spoken languages used around the world vary greatly in their vocabularies, grammars, and sound systems [1]. Some of this variation is the product of random drift and founder effects and is unpredictable once phylogenetic relatedness is taken into account. However, some of the variation may stem from languages adapting to different niches, that is, different ecological and sociodemographic environments in which they are learned and used [2]. For example, tonal languages appear to cluster in more humid environments [3], and grammatical complexity may covary with smaller speaker populations (reviewed in [4]). These observations and many others have led to a growing interest in understanding languages as flexible systems that adapt to cognitive [5], cultural [6], and geographical [7] niches.

Recent work by Aceves and Evans (A&E) [8] shines a spotlight on another aspect of linguistic variation, namely the efficiency with which different languages convey information. A&E assembled a corpus of parallel translations spanning hundreds of languages and multiple genres (e.g., Bibles, TED talks, movie subtitles, and parliamentary proceedings). They then computed information density for each language by encoding each word into its Huffman code (one of the first steps when compressing a file using a program such as gzip). The more bits it takes to encode a given text, the less its informational density.

One of A&E's key results is that information density is closely related to semantic density. A language is semantically dense to the extent that it has a smaller average semantic distance between words (computed using now-standard techniques for word embeddings). Other intriguing results include finding that more informationally dense languages have lower semantic breadth, a measure that A&E computed from transcribed conversations and Wikipedia articles. Conversations in informationally dense languages tend to proceed more quickly, but over a narrower semantic space.
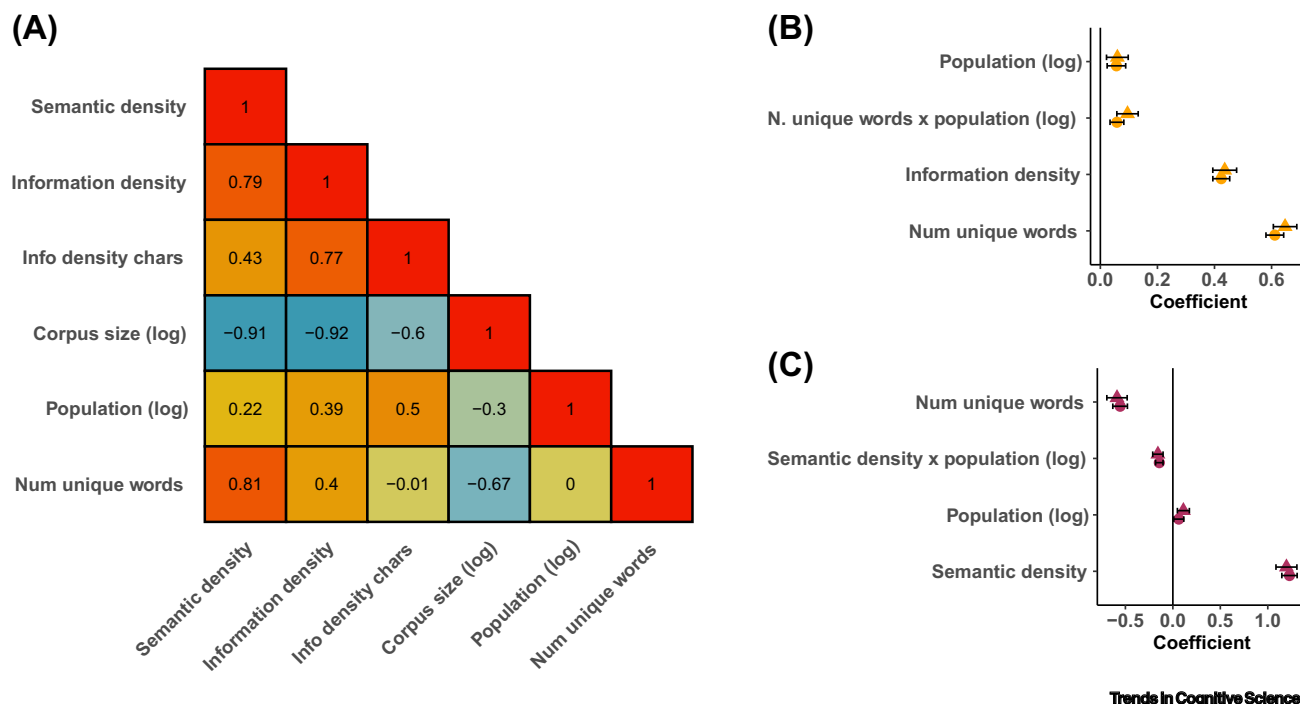
A&E's findings raise several questions. The close link ($r > 0.7$) between information density and semantic density makes one wonder whether it is caused by some third factor. The information density of a corpus computed using Huffman codes is very strongly ($r > 0.9$) related to the number of words in that corpus. This makes sense: if it takes a language more words to convey the 'same' message, then each word conveys less information. However, it is intriguing that taking fewer words to express a message (higher information density) is associated with higher semantic density. One could imagine the opposite, wherein words of a more informationally dense language span larger semantic distances. We can glean further insight into the possible causes of variation in semantic density, and its relationship to information density, by examining the role of vocabulary size across different languages. Predicting semantic density from information density and vocabulary size (the number of unique words in the New Testament corpus) shows that higher semantic density is strongly linked to vocabulary size. In fact, the latter is a better predictor than is information density (Figure 1). A&E's data also show an intriguing interaction: the larger the vocabulary size, the stronger the relationship between information density and semantic density.

A second question raised by A&E's findings is why some languages are more informationally and semantically dense in the first place. Is information density just a matter of chance, a product of a random process, such as drift? Or might it be the result of languages adapting to different functional pressures?

A&E focus on information and semantic density measures as key variables, showing that they remain predictive after controlling for other variables, such as population size, the geographical spread of each language, and several predictors related to climate in which the language is spoken. However, rather than potential confounds, associations with these variables can help us understand what gives rise to linguistic variation in the first place (Figure 1A). Consider, for example, population size, that is, the number of speakers that use a given language. Population size has been a focus in past work on linguistic diversity [2] for multiple reasons. Languages with small speaker populations tend to be learned in a 'tighter' social context, with speakers sharing more common ground and more likely to be native speakers. In contrast, languages with more speakers (including lingua francas and global languages, such as English) often comprise much more heterogeneous populations and speakers who learn the language as adults. A larger population also brings with it more opportunities for language contact and cross-cultural exchange.

This framing of population leads to two predictions. First, to the extent that greater

(A)

(B)

(C)

Trends in Cognitive Sciences

Figure 1. Information and semantic densities form a complex web of relationships. Whereas Aceves and Evans (A&E) focused on density measures [8], we illustrate some of the other important, mutually correlated factors that, along with density, drive language change. (A) First-order Spearman correlations among some variables of interest. Results from mixed-effects multiple regression showing standardized regression coefficients ± 95% confidence intervals (CIs) with by-family random effects predicting (B) semantic density and (C) information density. Number of unique words (vocabulary size) becomes negative once semantic density is included. Analyses include 801 languages from 89 language families. Results remain qualitatively similar (triangular markers) when we restrict the analysis to 268 'sister languages', a stricter control for phylogenetic relatedness. Full code for the visualization and underlying models is available at https://github.com/racdale/language-density-dynamics.

semantic density may arise from a series of linguistic innovations, larger populations should be associated with greater semantic densities. Second, children's word learning is facilitated by lower contextual surprisal [9], precisely what is offered by lower information density. Languages with smaller populations can be better optimized for being learned by children as a native language [2]. For such learners, lower information density becomes an asset because it acts to reduce surprisal. Revisiting A&E's open data confirms both predictions (Figure 1B,C).

Population also enters into some intriguing interactions. For example, when predicting information density, we observe a negative interaction between population and semantic density: the link between semantic and information density is stronger for languages with larger populations. A possibility

(although a speculative one) is that this relationship arises from larger speech communities having more opportunities for linguistic innovation. Some of the innovations will lead to the coining of words that efficiently fill communicative needs in ways that increase both information and semantic density. This prediction can be experimentally tested using artificial language learning experiments.

A&E's study is an exciting jumping-off point for further investigating mechanisms that drive variation in information structure across languages. Large-scale investigations such as these have their sharpest epistemic edge when refuting claims of strict universalism in language structure, acquisition, and use (e.g., [10] for discussion). We suspect that measures such as density and breadth are among an ensemble of factors in a causal web, untangling

which will require testing theory-driven predictions using both correlational and experimental methods.

Declaration of interests

None declared by authors.

[1]Department of Psychology, University of Wisconsin-Madison, Madison, WI, USA
[2]Department of Language Science and Technology, Saarland University, Saarbrücken, Germany
[3]Department of Communication, UCLA, Los Angeles, CA, USA

*Correspondence:
lupyan@wisc.edu (G. Lupyan), pablock@lst.uni-saarland.de (P. Contreras Kallens), and rdale@ucla.edu (R. Dale).

References

1. Evans, N. and Levinson, S.C. (2009) The myth of language universals: language diversity and its importance for cognitive science. *Behav. Brain Sci.* 32, 429–448

2. Lupyan, G. and Dale, R. (2016) Why are there different languages? The role of adaptation in linguistic diversity. *Trends Cogn. Sci.* 20, 649–660

3. Everett, C. *et al.* (2015) Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. *Proc. Natl. Acad. Sci. U. S. A.* 112, 1322–1327

4. Five Graces Group *et al.* (2009) Language is a complex adaptive system: position paper. *Lang. Learn.* 59, 1–26

5. Christiansen, M.H. and Chater, N. (2008) Language as shaped by the brain. *Behav. Brain Sci.* 31, 489–509

6. Trudgill, P. (2011) *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*, Oxford University Press

7. Pacheco Coelho, M.T. *et al.* (2019) Drivers of geographical patterns of North American language diversity. *Proc. R. Soc. B* 286, 20190242

8. Aceves, P. and Evans, J.A. (2024) Human languages with greater information density have higher communication speed but lower conversation breadth. *Nat. Hum. Behav.*, Published online February 16, 2024. https://doi.org/10.1038/s41562-024-01815-w

9. Portelance, E. *et al.* (2023) Predicting age of acquisition for children's early vocabulary in five languages using language model surprisal. *Cogn. Sci.* 47, e13334

10. Christiansen, M.H. *et al.* (2022) Toward a comparative approach to language acquisition. *Curr. Dir. Psychol. Sci.* 31, 131–138