Word lengths are not optimized for efficient communication

Amanda Doucette[1] **Timothy J. O'Donnell**[1,2] **Morgan Sonderegger**[1]
[1]McGill University [2]Canada CIFAR AI Chair, Mila
amanda.doucette@mail.mcgill.ca
timothy.odonnell@mcgill.ca morgan.sonderegger@mcgill.ca

Zipf's law of abbreviation (Zipf, 1935) describes a negative relationship between word length and frequency that appears to be a linguistic universal (e.g. Bentz and Ferrer-i Cancho, 2016; Petrini et al., 2023). However, this has been called into question: Piantadosi et al. (2011) suggests that Zipf's law is better characterized as a relationship between average information content (or *surprisal*) and word length, rather than between frequency and word length. Average information content, as defined by Piantadosi et al. (2011), is the average probability of a word occurring in a particular context. In 11 languages, they show that average information content more strongly correlates with word length than frequency correlates with word length.

While these correlations show that there is indeed a relationship between surprisal and word length, they do not show that words are short *because* they are highly informative. In this study, we directly address the question of whether Zipf's law is better characterized as a causal relationship between word length and frequency, between word length and average surprisal, or if there is a causal relationship in some languages, but only a correlational relationship in others. Following Doucette et al. (2025), we apply a causal discovery algorithm (Fast Causal Inference, Spirtes et al., 1995) to 12 languages selected from the LexiPhon dataset (Doucette et al., 2026) to examine causal relationships between word length (in phones and in syllables), frequency, surprisal, and two additional lexical variables that are known to correlate with length and frequency: morphological irregularity and phonotactic complexity.

Surprisal is defined as the negative log probability of a word w_t given its previous context $w_{<t}$: $-\log p(w_t \mid w_{<t})$. We calculate average surprisal for each word using occurrences of that word in Wikipedia, following the method described by Pimentel et al. (2023) and using the XLM-RoBERTa language model (Conneau et al., 2020) to estimate word probability in context. For morphological irregularity and phonotactic complexity, we use data provided by Doucette et al. (2025). Number of syllables is estimated by counting the number of vowels in each word.

Edge	# Languages
FR $\longleftrightarrow$ SRP	4
PC $\longleftrightarrow$ WL	8
SRP $\longleftrightarrow$ WL	6
SYL $\longleftrightarrow$ PC	7
SYL $\longrightarrow$ WL	6
FR - - - MI	6
FR - - - PC	9
FR - - - SYL	8
FR - - - WL	11
MI - - - PC	4
MI - - - WL	8
SRP - - - MI	9
SRP - - - PC	10
SYL - - - MI	9
SYL - - - SRP	6

Table 1: Most likely edge types with number of languages out of 12 where edge appears in most likely graph. No edge is represented by - - - . WL: length in phones; SYL: length in syllables; FR: frequency; SRP: average surprisal; PC: phonotactic complexity; MI: morphological irregularity.

We use a bootstrapping procedure to resample the data for each language 3000 times, then apply the Fast Causal Inference algorithm to each sample. The algorithm outputs a Partial Ancestral Graph (PAG), a type of causal graph that can represent both causal relationships between the measured variables and the presence of possible unmeasured confounding variables. A unidirectional arrow in the graph represents a direct causal relationship, while a bidirectional arrow represents a relationship where an unmeasured confounding variable is a common cause. A circle rather than an arrowhead

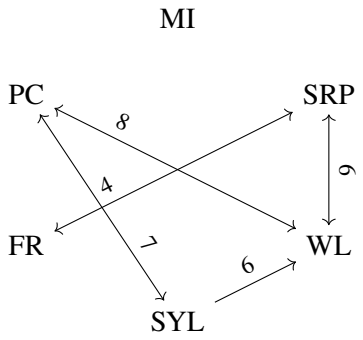

Figure 1: Most likely PAG across 12 languages, annotated with number of languages where edge occurs.

represents uncertainty in the causal relationship. For each language, we aggregate the 3000 graphs by taking the most commonly occurring edge type between each pair of variables.

We find that each language has a unique causal graph, suggesting that there is no cross-linguistic universal causal structure in the lexicon. Even typologically related languages such as English and German do not share the same causal structure. Table 1 shows the most likely edge types to occur across the 12 languages, and is represented graphically in Figure 1.

Perhaps most striking is that 11 of 12 languages have no edge between word length in phones and frequency, which would be expected under a causal interpretation of Zipf's law of abbreviation. We also see that 8 of 12 languages have no edge between number of syllables and frequency, another possible instantiation of Zipf's law. This suggests that Zipf's law is not the result of a universal causal relationship between length and frequency. We also see that the most frequent edge between average surprisal and word length in phones is SRP $\longleftrightarrow$ WL in 6 of 12 languages, and the most frequent edge between average surprisal and number of syllables is no edge, in 6 of 12 languages. This suggests that the causal relationship between word length and average surprisal is even less consistent across languages than the causal relationship between word length and frequency.

In other words, we see little evidence that Zipf's law of abbreviation is a universal causal relationship between word length and *either* frequency or average surprisal. Instead, the well-established universal correlation between length and frequency is the result of different causal mechanisms that can vary across languages. In the case of many languages in our data set, that causal mechanism involves variables outside of word length, frequency, average surprisal, phonotactic complexity, and morphological irregularity. Whether there is a stronger correlation between length and frequency or between length and average surprisal – as has been debated by Piantadosi et al. (2011), Levshina (2022), Pimentel et al. (2023), among others – has little bearing on the underlying causal relationships.

References

Christian Bentz and Ramon Ferrer-i Cancho. 2016. Zipf's law of abbreviation as a language universal. In *Proc. of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proc. 58th ACL*, pages 8440–8451.

Amanda Doucette, Timothy J. O'Donnell, and Morgan Sonderegger. 2025. Unzipping the causality of Zipf's law and other lexical trade-offs. In *Proc. of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 66–76.

Amanda Doucette, Timothy J. O'Donnell, and Morgan Sonderegger. 2026. LexiPhon: A collection of phonetically transcribed lexicons from Wikipedia. In *Proc. LREC*. To appear.

Natalia Levshina. 2022. Frequency, informativity and word length: Insights from typologically diverse corpora. *Entropy*, 24(2).

Sonia Petrini, Antoni Casas-i Muñoz, Jordi Cluet-i Martinell, Mengxue Wang, Christian Bentz, and Ramon Ferrer-i Cancho. 2023. Direct and indirect evidence of compression of word lengths. Zipf's law of abbreviation revisited. *Glottometrics*, 54:58–87.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *PNAS*, 108(9):3526–3529.

Tiago Pimentel, Clara Meister, Ethan Wilcox, Kyle Mahowald, and Ryan Cotterell. 2023. Revisiting the optimality of word lengths. In *Proc. EMNLP*, pages 2240–2255.

Peter Spirtes, Christopher Meek, and Thomas Richardson. 1995. Causal inference in the presence of latent variables and selection bias. In *Proc. of the 11th Conference on Uncertainty in Artificial Intelligence*, page 499–506.

George Kingsley Zipf. 1935. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Houghton Mifflin.