The Laplacian spectrum of language: experiments with linguistic families from the Americas

– DOI: https://doi.org/10.1051/2025405

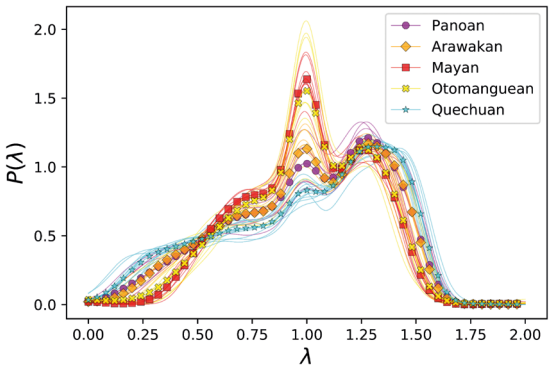
What can physics offer to the study of endangered languages? In a world where only a small fraction of the approximately 7,000 spoken languages are represented in the current wave of artificial intelligence—through tools like machine translation, voice assistants, or large language models—most Indigenous languages, especially those of South America, remain digitally invisible.

Their absence is not just a technical limitation; it reflects deep inequalities in how knowledge, culture, and language are treated in the digital age.

This work explores how tools from statistical mechanics, and in particular spectral graph theory, can contribute to the computational representation of low-resource languages. Using parallel translations of the Universal Declaration of Human Rights, word co-occurrence graphs are constructed for 85 languages of the Americas. In each graph, nodes represent unique words, and edges are created between words that appear within a short distance of each other in the text. The edges are weighted according to how often the word pairs appear together. These graphs offer a simple, language-agnostic way of encoding the structure of the text.

To analyze these graphs, a normalized version of the Laplacian matrix is computed. This matrix captures the graph's internal structure by describing how each word is connected to its neighbors. The eigenvalues of this matrix—known as the Laplacian spectrum—form a statistical signature that summarizes key features of the graph. In physics, such spectral profiles are widely used to study the dynamics and geometry of complex systems. Here, they are repurposed to describe patterns in human language.

The analysis shows that different linguistic families—such as Quechuan, Arawakan, Mayan, and Otomanguean—produce distinct spectral patterns. These patterns are closely related to known linguistic properties. For instance, the frequency of one particular eigenvalue (commonly associated with node duplication in graph theory) is negatively correlated with word entropy, a measure of lexical diversity. Languages with richer morphological systems tend to show higher entropy and lower frequencies of this eigenvalue, suggesting that meaningful structural differences are reflected in the spectra.



▲ Laplacian spectra of linguistic families from the Americas. The figure displays the average Laplacian spectra of five major linguistic families spoken in South America: Panoan, Arawakan, Mayan, Otomanguean, and Quechuan. Each curve is constructed by averaging the spectral profiles—sets of eigenvalues from the normalized Laplacian matrix—across all languages within a family. These spectra are derived from word co-occurrence graphs based on translations of the Universal Declaration of Human Rights, using a graph radius of two. Peaks at specific eigenvalues suggest the presence of structural regularities shared within linguistic families. These spectral shapes offer a compact, data-light way to detect typological differences and similarities, especially valuable in the context of digitally marginalized languages. By translating linguistic variation into spectral form, the figure illustrates how physical tools can reveal hidden dimensions of cultural diversity and contribute to the computational visibility of endangered languages.

One key advantage of this approach is that it does not require large datasets, linguistic annotations, or deep language-specific knowledge. Instead, it offers a lightweight, interpretable, and reproducible method to study languages with limited digital presence. This makes it especially suitable for endangered and marginalized languages that are often left out of computational linguistics and AI development.

Beyond its technical contribution, this study encourages a broader reflection on how physics can support the digital humanities. Graph-based methods, long used in physics to model interactions and flows, can also reveal the hidden geometries of language. The Laplacian spectrum becomes not just a mathematical construct, but a way to detect structure, similarity, and diversity across languages—even in minimal texts.

Ultimately, applying classical tools from physics to cultural data is also a gesture of inclusion. It points to new roles for physics in building digital infrastructures that are more equitable and representative. By making small, endangered languages visible in a computational sense, this work contributes to a more inclusive vision of how data—and science—can serve cultural diversity. ■

Reference

[1] J. Vera Zúñiga, The Laplacian spectrum of language: Experiments with linguistic families from the Americas, *EPL* **150** (2025) 22001. https://doi.org/10.1209/0295-5075/adb450