LETTER • OPEN ACCESS

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

To cite this article: Javier Vera Zúñiga 2025 *EPL* **150** 22001

A Letters Journal Exploring the Frontiers of Physics

April 2025

EPL, **150** (2025) 22001
doi: 10.1209/0295-5075/adb450

www.epljournal.org

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

Javier Vera Zúñiga[(a)]

Departamento de Humanidades, Pontificia Universidad Católica del Perú - Lima, Peru

received 16 October 2024; accepted in final form 10 February 2025
published online 28 April 2025

Abstract – To what extent would statistical mechanics approaches help to represent languages from the Americas? Is it possible to extract useful information about the relationships between these languages? This work studies a graph-based approach to extract information from text corpora of languages of the Americas. Each language is viewed as the set of eigenvalues obtained from the Laplacian matrix of co-occurrence graphs. The results suggest that our graph-based feature extraction technique is partly comparable to the knowledge contained in typological databases. We argue that our approach might propose a solution to the lack of textual resources for low-resource languages.

Introduction. – To what extent would statistical mechanics approaches help to represent languages from the Americas? Is it possible to extract information about the relationships between these languages? This paper proposes a graph-based approach to (partly) answer the previous questions. To do this, we deal with statistical mechanics approaches to low-resource languages (LRLs), particularly from the Americas. To face with LRLs, we refer them as *resource scarce*, *less privileged* or *endangered* [1]. Crucially, most of the LRLs are indigenous languages in which the cultures and knowledge systems to which they belong are put at risk.

The motivation for classifying and analyzing these languages goes beyond theoretical interest. From a practical standpoint, understanding the structure and relationships of indigenous languages can contribute to their preservation by enabling the development of computational tools, such as automatic translation systems, digital dictionaries, and educational platforms [2]. These tools can support the revitalization of languages at risk of disappearing, which, in turn, safeguards the cultural heritage and knowledge systems embedded in these languages [3]. Moreover, classifying languages can help document and formalize linguistic patterns, providing a foundation for future linguistic and anthropological research [4,5].

With this in mind, our main goal is to propose a statistical mechanics approach to the extraction of typologically valuable features from text corpora [6]. This work follows thus a graph-based approach to describe the organization of languages of the Americas. Crucially, inspired by [7], we strongly believe that technological and computational approaches to LRLs can have a positive social impact for the communities which depend on these languages.

To face our computational approach to LRLs, we used the Laplacian spectrum [8–10] of graphs representing textual data. Within this framework, the authors of [10] have revealed global properties of neural networks, describing the connections between neural elements. The Laplacian spectrum is formed by the set of eigenvalues obtained from the *normalized Laplacian matrix* of a graph. Remarkably, this set of numbers captures not only the global properties of a graph, but also local structures that are produced by graph changes (like motif or node duplication) [11]. Therefore, the Laplacian spectrum allows us to propose a simple way to represent and extract information from text corpora.

Recent work on statistical mechanics approaches to language has remarked the fact that language can be modeled by graphs [5,12–14]. In this paper, feature representation of languages [15–18] is associated to co-occurrence graphs whose edges capture thus inter-word relationships [19,20]. An interesting related work in this line is [21], in which co-occurrence representation of languages allowed the extraction of several graph-mining measures in order to apply machine learning models.

The remaining of the article details the graph-based approach to feature extraction for languages of the Americas.

[(a)]E-mail: jveraz@pucp.edu.pe (corresponding author)

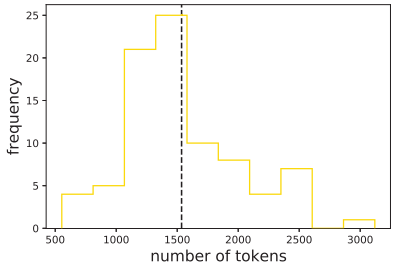

Fig. 1: Basic description of the parallel corpus of languages of the Americas. The figure displays the histogram of the number of tokens for each text of the *UDHR* corpus. We consider the 85 languages appearing in both *UDHR* and the *WALS* typological database. The black dotted line indicates the average value.

Table 1: Basic description of the linguistic families from the Americas (based on [22]). The languages studied in this paper are those appearing in both *UDHR* and the *WALS* typological database [23].

Linguistic family	Glottocode	Languages
Jivaroan	jiva1245	3
Panoan	pano1256	6
Arawakan	araw1281	8
Mayan	maya1287	8
Barbacoan	barb1265	3
Otomanguean	otom1299	8
Algic	algi1248	3
Quechuan	quec1387	13
Other families		39
Total		85

We organize the discussion in three sections. The next section "Materials and methods" describes language data and our graph-based approach to linguistic complexity. Section "Results" describes and illustrates the main results. Section "Discussion" summarizes our work and restates the key challenges of our approach to the computational representation of LRLs.

Materials and methods. –

Universal declaration of human rights corpus UDHR. To extract graph-based features, we need a comparable text corpora across many languages, in which text content is constant in order to avoid style or genre distortions. To control this constraint across languages, we use a parallel corpus of the *universal declaration of human rights*[1] (*UDHR*). A *word type* is defined here as a unique string delimited by white spaces. A *word token* is then any repetition of a word type. Details about the *UDHR* corpus are shown in fig. 1 and table 1.

The World Atlas of World Languages (WALS). Typological information about languages of the Americas is obtained from *WALS* [23]. The World Atlas of World Languages (*WALS*) is a large dataset of structural properties of thousands of languages obtained from descriptive grammars. Using this dataset, each language can be represented as a list of feature values. To overcome the sparsity of *WALS*, particularly in relation to low-resource languages of the Americas, we used *lang2vec* syntactic vectors comprising 103 features [2][2]. The languages studied in this paper are those appearing in both *UDHR* and the *WALS* typological database [23] (through *lang2vec* features). We used *iso-639-3* codes to find the intersection between languages appearing both in the *UDHR* corpus and in [23].

Information-theoretic entropy. As suggested by the seminal works of Shannon [24], the choice associated with words is a key property of human language. At the center of information theory, there is a precise quantity of the average amount of choice associated with words: the *word entropy*.

Let T be a text formed by word types taken from the set W_t. In probabilistic terms, word type probabilities are distributed according to $p(w)$, $w \in W_t$. The *average amount of choice of word types* (or simply the *entropy*) reads [25]

$$H = -\sum_{w \in W_t} p(w) \log(p(w)). \quad (1)$$

Using several corpora and tackling some problems of word entropy estimation, [25] provided a public database of entropy values for 1259 languages. Since all entropy estimators are strongly correlated, for our experiments we used the *NSB* estimator [26].

For the large sample of languages of the world, the authors of [25] found high- and low-entropy areas. A key example is languages of the Andean region of South America. These languages all have high word entropies. This can be interpreted as a result of their high morphological complexity, arising from a large *type token* ratio [27].

Basic concepts on (spectral) graph theory. To extract the Laplacian spectrum for each language (represented by its respective translation of the *UDHR*), we need to introduce some concepts of (spectral) graph theory. We consider an undirected and weighted graph $G = (V, E, W_E)$, where V represents the set of nodes of size n, E is the set of edges and W_E is the set of weights. In our approach, the set V represents word types for a language, while E is formed by pairs of word types at radius lower than or equal to 3. The distance between nodes is measured in word types: consecutive words are at distance (or *radius*) 1, words separated by just one word are at distance (or *radius*) 2, and so on. The *weight* $w(uv) \in W_E$ associated to the edge $uv \in E$ counts the number of appearances of the (possibly long-distance) bigram uv. The *neighborhood* of the node $u \in V$ is the set $V_u = \{v \in V : uv \in E\}$. The (weighted) *degree* of the node $u \in V$ is simply the sum of the weights joining u with nodes in V_u.

Spectral graph theory is mainly focused on discovering graph properties arising from the eigenvalues of the

[1] `https://www.unicode.org/udhr/index.html`.

[2] `https://github.com/antonisa/lang2vec`.

matrices associated to graphs [28], such as the *adjacency matrix* and the *Laplacian matrix*.

The *normalized Laplacian matrix* is defined by the relation $\mathcal{L} = I - D^{-1/2}AD^{-1/2}$, where A is the adjacency matrix; I is the $|V| \times |V|$ identity matrix; and D is a diagonal matrix whose entries are (possibly weighted) node degrees. The Laplacian spectrum of the graph G is the collection of all solutions λ (the *eigenvalues*), for which there exist nonzero vectors u (the corresponding *eigenvectors*) satisfying the equation $\mathcal{L}u = \lambda u$. An interesting property is that the largest eigenvalue is always equal or smaller than 2, labeling the range of eigenvalues as $0 \leqslant \lambda_1 \leqslant \ldots \leqslant \lambda_n \leqslant 2$ [28]. With this, a language can be represented as the list of eigenvalues of $\mathcal{L}$, which has been constructed using the respective translation of the *UDHR* corpus.

Graph construction and implementation details. Basic management of typological data was made using *pandas* [29,30]. For text preprocessing (whitespace tokenization, punctuation removal and conversion to lower case), we used *NLTK* [31]. Graph-theoretic techniques were made using *NetworkX* [32] and *NumPy* [33].

Each language is represented first by a co-occurrence graph G, and then by the *normalized Laplacian matrix* $\mathcal{L}$ associated to G. The graph G was built along the following steps:

Step 1. Identify the set of sentences.

Step 2. Preprocess each sentence by whitespace tokenization, punctuation removal and conversion to lower case.

Step 3. Define the set of word types W_t of the entire text.

Step 4. Through an iterative process, inspect each sentence in order to find word type co-occurrences at distances one, two and three (based on the fact that dependence relationships occur in general at small distances [34]). Each new co-occurrence between pairs of word types from W_t defines an edge of the graph. Repetitions of bigrams increase the weight of the respective edge.

To enhance the graph-building process, alternative configurations were considered. For example, instead of relying solely on co-occurrence frequencies, a semantic-based approach could link words with similar meanings or contextual usage. However, these alternatives often require additional linguistic resources, such as pre-trained word embeddings or annotated corpora, which are typically unavailable for low-resource languages [19,21]. By focusing on co-occurrence graphs, this study prioritizes a scalable and data-light methodology tailored to the challenges of working with low-resource languages.

For simplicity, our results focus on the Laplacian spectra obtained from co-occurrence graphs of radius one, two and three. With the co-occurrence graph G, we calculate the set of eigenvalues of the *normalized Laplacian matrix* $\mathcal{L}$. This set is arranged by increasing size: $\Lambda = \{0, \lambda_1, \ldots, \lambda_n\}$.

Calculations with Laplacian spectra: distances and nearest-neighbor graphs. An important issue concerns the comparison of eigenvalue sets of different size. To face this problem, spectral plots were obtained from a smoothed eigenvalue distribution Λ^*, consisting of eigenvalue frequencies estimated by a *kernel density estimation* with bandwidth 0.065 [35]. To compare language representations of the same size, a discrete smoothed spectrum was used, in which Λ^* had 1000 predicted values. All calculations were made using *SciPy* [36].

Using predicted Laplacian spectrum distributions, we define a simple *distance* measure between languages. Consider two smoothed eigenvalue distributions $\Lambda^*_A = \{\lambda^A_1, \ldots, \lambda^A_{1000}\}$ and $\Lambda^*_B = \{\lambda^B_1, \ldots, \lambda^B_{1000}\}$, representing respectively languages A and B. The distance between A and B reads

$$d(A,B) = \frac{1}{1000} \sum_{\lambda^A_i \in \Lambda^*_A, \lambda^B_i \in \Lambda^*_B} |\lambda^A_i - \lambda^B_i|. \qquad (2)$$

With a distance measure, we define a matrix of distances between languages D. There are several ways to transform a *distance matrix* D into a graph. A nice (and simple) graph-theoretic way relies on a *similarity graph* $G^D = (V, E)$. Each language L_i is understood as a vertex $v_i \in V$. The vertex (language) v_i is joined by an edge ($\in E$) with the vertex v_j if L_j is among the k nearest neighbors of L_i. Put differently, from the *distance matrix* D we rank the languages of the row (or column) i by increasing distance d. The vertex v_i is joined thus to the k closest languages (excluding the language i itself).

Justification for using the Laplacian spectrum. The choice of the Laplacian spectrum as a key metric in this study is motivated by its ability to capture both global and local structural properties of graphs. Unlike traditional network metrics (*e.g.*, degree distribution, clustering coefficient) or neural embeddings, the Laplacian spectrum provides a comprehensive representation of graph topology that is sensitive to subtle changes in structure. For example, it reflects variations in connectivity patterns, such as motif duplications or node distributions, which are critical to modeling linguistic features [11,28].

Additionally, the Laplacian spectrum is computationally efficient for small datasets, making it particularly suitable for low-resource languages where data availability is limited. This approach contrasts with methods that require extensive semantic annotation or pre-trained embeddings, which are often unavailable for such languages [19]. Furthermore, spectral representation allows direct comparisons between languages through eigenvalue distributions, allowing the identification of typological patterns that might be obscured by other methods [34].

Our use of the Laplacian spectrum also aligns with recent advances in spectral graph theory, which emphasize

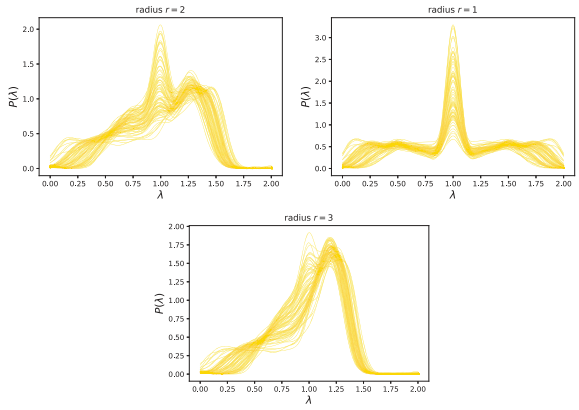

Fig. 2: Laplacian spectra of languages from the Americas. The figure shows the spectrum of the co-occurrence graphs for 85 languages from the Americas. Each panel indicates the spectra for co-occurrence graphs based on radius one, two or three. Laplacian spectra showed differences at specific eigenvalues, suggesting that a vector representation of language structure can be recovered.

its utility to uncover hidden structural properties in complex systems, including language networks [5,12]. By focusing on eigenvalues, we aim to provide a robust and interpretable representation of linguistic data that can bridge the gap between graph theory and linguistic typology.

Results. – In what follows, we first give a qualitative description of the Laplacian spectrum of the languages from the Americas. To understand the *shape* of the Laplacian spectrum, we provide a simple comparison between the frequency of the eigenvalue λ_1 and word entropy values. Second, we give simple observations about the relationships established by linguistic families from the Americas. Then, we evaluate how much information about typological features is recovered by the Laplacian spectrum approach. Finally, based on distance measures we applied some spectral clustering techniques over nearest-neighbors graphs representing languages and groups of languages from the Americas.

Laplacian spectrum for languages of the Americas. The Laplacian spectra of the languages from the Americas revealed several interesting properties, as shown in fig. 2. First, there is a clear qualitative influence of the radius (one, two or three) on the shape of the Laplacian spectra. Further work should study such differences. Second, despite the mentioned differences all spectra displayed a notorious peak at eigenvalue one.

Particularly for radius two, all the Laplacian spectra showed a similar distribution in which the lowest and the largest eigenvalues are, respectively, 0 and 1.75. All distributions showed peaks at specific eigenvalues: 0.25, 0.75, 1, 1.25 and 1.5. The existence of these peaks suggest underlying structural properties of languages (and linguistic families) from the Americas.

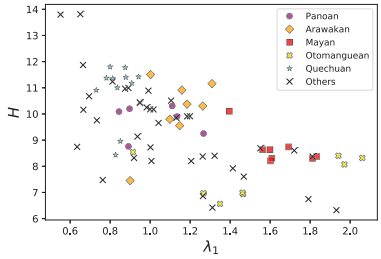

Fig. 3: Frequency of eigenvalue one λ_1 *vs.* word entropy H. According to our corpus of 85 languages from the Americas, largest linguistic families (*Panoan*, *Arawakan*, *Mayan*, *Otomanguean* and *Quechuan* languages) are plotted in the two-dimensional space formed by the frequency of the eigenvalue one and word entropy. Other languages are indicated with black $\times$. Results are based on co-occurrence graphs with radius two.

To interpret the appearance of peaks at eigenvalue one, we first remark that node duplication, in which one new node is introduced to a graph (with the same connectivity pattern of the duplicated node), results in an increase in the eigenvalue $\lambda = 1$ [8]. As shown in fig. 3 (for radius two), the frequency of eigenvalue one λ_1 is negatively correlated with word entropy H.

The non-parametric Spearman rank correlation gives -0.624 (radius one); -0.645 (radius two); and -0.610 (radius three) ($p < 0.001$ level). This suggested an interesting fact: low word entropy linguistic families (*Maya* and *Otomangue*) exhibited at the same time large frequencies of λ_1. This might indicate higher levels of node duplication in languages of low word entropy.

Relationships between linguistic families. The Laplacian spectra of linguistic families from the Americas are shown in fig. 4 (top). As previously suggested, linguistic families differ at specific eigenvalue peaks. In qualitative terms, *Mayan* and *Otomanguean* languages seem to be closer regarding *Panoan*, *Arawakan* and *Quechuan* languages.

As expected, linguistic subfamilies tend to be grouped, as shown in fig. 4 (bottom). Given that similar Laplacian spectra reflect similar graph properties, we might notice that relationships within linguistic families (for example, between *Quechua I* and *Quechua II*) are recovered by their spectra distributions.

Comparisons with typological information. In order to compare Laplacian spectrum distributions with typological data, we applied the *t-SNE* [37] dimensionality reduction technique (using the *scikit-learn* implementation [38]). This technique is particularly well designed for embedding high-dimensional data. In our case, we reduced the dimensionality of the two main sources of data to a two-dimensional space: 10000 eigenvalues of the Laplacian spectra; and 103 typological features. As shown in fig. 5, some previous observations about the organization of linguistic families of the Americas can be recovered from the

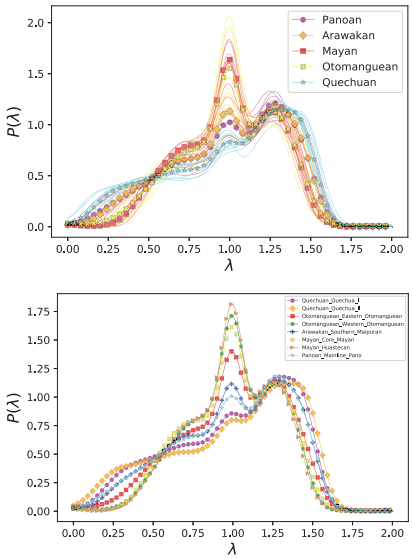

Fig. 4: Laplacian spectra of linguistic families from the Americas. The top panel shows the Laplacian spectra of a number of linguistic families from the Americas (*Panoan*, *Arawakan*, *Mayan*, *Otomanguean* and *Quechuan* languages). The bottom panel shows the Laplacian spectra of the corresponding linguistic subfamilies (according to the available data). Languages belonging to a specific linguistic family are represented by colors. A group of languages (linguistic family or subfamily) is represented by the average of the Laplacian spectra of its associated languages. Results are based on co-occurrence graphs with radius two.

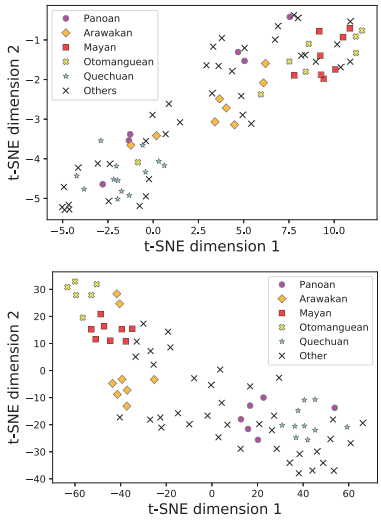

Fig. 5: Two-dimensional representation of Laplacian spectrum distributions and typological dara using *t-SNE*. According to our corpus of 85 languages from the Americas, largest linguistic families (*Panoan*, *Arawakan*, *Mayan*, *Otomanguean* and *Quechuan* languages) are plotted in the two-dimensional space formed by the embedding of 10000 eigenvalues of the Laplacian spectra (top); and 103 typological features (bottom). Other languages are indicated with black ×. Results are based on co-occurrence graphs with radius two.

two-dimensional embedding of Laplacian spectra. Despite the fact that embedding high-dimensional datasets into low-dimensional spaces supposes some distortions, *Mayan* and *Otomanguean* languages can be grouped as a separated linguistic area.

Confirming visual observations on two-dimensional embeddings, we give a quantitative approach to the formation of language areas based on a simple classification task for linguistic areas: *Mayan* and *Otomanguean* languages *vs.* *Panoan*, *Arawakan* and *Quechuan* languages. We trained two *Random Forest* [38] classifiers combining 100 individual trees: a 2-way classifier that takes 103 typological features taken from *lang2vec*; and a 2-way classifier using the smoothed eigenvalue distribution Λ^* to represent each language. The two classifiers obtain F1-measures of 96.96 $\pm$ 0.08 and 87.27 $\pm$ 0.10 (3-fold cross-validation), respectively. This result suggests that the task of separating into two language areas can be solved with high accuracy for the Laplacian spectrum representation of languages.

Clustering of related languages. To extract information from distance matrices, which define pairwise similarities between pairs or group of languages, we used two inter-related techniques. First, we clustered related languages using a *heatmap*, as shown in fig. 6. In this representation, each linguistic subfamily cell is represented by a color. Darker colors indicate increasing distances between pair of linguistic subfamilies. Columns and rows were rearranged by unsupervised *agglomerative clustering* [32,39]. Unlike the classification results, we remark thus that no label has been assigned to the linguistic subfamilies. A first important observation relies on the number of clusters. Dendrograms illustrated how each cluster is composed by small groups of languages. For both panels, *Mayan* and *Otomanguean* languages are clustered together. As expected, for both panels *Quechuan* languages behave as a linguistic unit. By contrast, the two sources of data (Laplacian spectra distributions and typological data) differ regarding *Arawakan* languages. Further work should study this observation.

Discussion. –

Laplacian spectra of low-resource languages from the Americas. In this paper, we described the Laplacian spectra of languages from the Americas, extracting typologically valuable information from parallel corpora. By contrasting Laplacian spectra with syntactic features (from *lang2vec*) and word entropy values, we demonstrated two key findings: linguistic families from the Americas can be represented by their Laplacian spectra, and there is a notable relationship between the frequency of the eigenvalue one λ_1 and word entropy values.

Digital representation of low-resource languages. Digital representations of low-resource languages are crucial in addressing inequalities embedded in data scarcity.

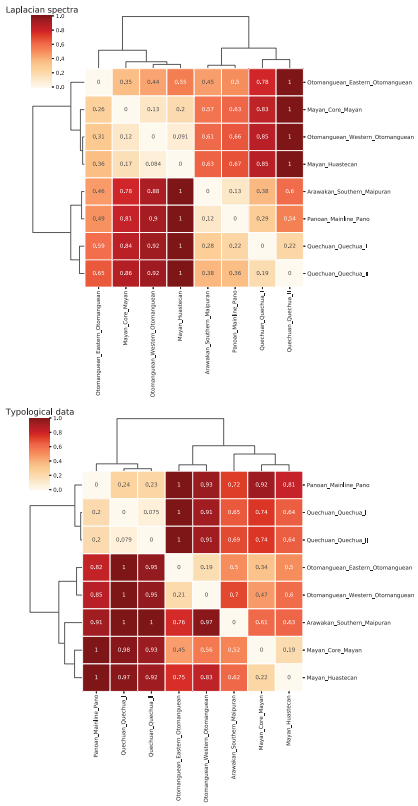

Fig. 6: Heatmaps representing Laplacian spectrum distributions and typological data. For smoothed Laplacian spectrum distributions (top) and *lang2vec* typological data (bottom), linguistic subfamily cells were rearranged using agglomerative clustering. Each subfamily is represented as the mean vector of its languages. Distances between Laplacian spectra are calculated by eq. (2), while distances between typological vector are calculated by the cosine between their representations. Cell numbers indicate normalized distances.

Hidden within this scarcity are endangered cultures and knowledge systems. While privileged languages like *English* benefit from advanced computational tools, low-resource languages often lack such advantages. Graph-based representations, as proposed here, require minimal data compared to modern neural network approaches, making them particularly suitable for addressing this gap.

Interdisciplinary and computational approaches to typology. Our approach aligns with Bender's critique of language independence in NLP systems [40], emphasizing the need for linguistic knowledge. Computational methods for analyzing linguistic structures bridge NLP, typology, computational linguistics, and statistical mechanics, fostering interdisciplinary collaboration to address variations in linguistic structure.

Comparison with other classification efforts. The proposed graph-based method provides a distinct perspective compared to recent studies utilizing parallel corpora or entropy measures [25,27]. While parallel corpora approaches excel for resource-rich languages, they are less effective for low-resource languages. Similarly, entropy measures capture probabilistic structures but lack the topological granularity provided by graph-based representations. By leveraging the Laplacian spectrum, our approach integrates global and local structures, offering complementary insights into linguistic typology.

Connection to linguistic complexity. Our findings contribute to debates on linguistic complexity, linking spectral graph metrics with established measures like entropy. High entropy values, indicative of morphological richness [25], correlate with eigenvalue distributions, suggesting that spectral properties serve as proxies for linguistic complexity. The observed peaks in the Laplacian spectrum reflect structural redundancies and variances in linguistic graphs, corresponding to morphological or syntactic features inherent to specific language families. This integration of graph-theoretic and information-theoretic approaches underscores the value of interdisciplinary methods in advancing linguistic typology and analyzing low-resource languages.

Data availability statement: The data that support the findings of this study are openly available at the following URL/DOI: `http://unicode.org/udhr/index.html`.

REFERENCES

[1] Magueresse A., Carles V. and Heetderks E., *Low-resource languages: A review of past work and future challenges*, arXiv:2006.07264 (2020).

[2] Littell P., Mortensen D. R., Lin K., Kairis K., Turner C. and Levin L., *URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors*, in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics:* Vol. **2**, *Short Papers* (Association for Computational Linguistics, Valencia, Spain) 2017 pp. 8–14, `https://www.aclweb.org/anthology/E17-2002`.

[3] Bentz C., *Adaptive Languages* (De Gruyter Mouton, Berlin, Boston) 2018, `https://doi.org/10.1515/9783110560107`.

[4] i Cancho R. F., Łukasz Debowski and del Prado Martín F. M., *J. Stat. Mech.: Theory Exp.*, **2013** (2013) L07001.

[5] Cong J. and Liu H., *Phys. Life Rev.*, **11** (2014) 598.

[6] Ponti E. M., O'Horan H., Berzak Y., Vulić I., Reichart R., Poibeau T., Shutova E. and Korhonen A., *Comput. Linguist.*, **45** (2019) 559.

[7] Mager M., Gutierrez-Vasques X., Sierra G. and Meza-Ruiz I., *Challenges of language technologies for the indigenous languages of the americas*, in *Proceedings of the 27th International Conference on Computational Linguistics* (Association for Computational Linguistics) 2018, pp. 55–69.

[8] Banerjee A. and Jost J., *Linear Algebra Appl.*, **428** (2008) 3015.

[9] Banerjee A. and Jost J., *Discrete Appl. Math.*, **157** (2009) 2425.

[10] de Lange S., de Reus M. and Van Den Heuvel M., *Front. Comput. Neurosci.*, **7** (2014) 189.

[11] Banerjee A., *Biosystems*, **107** (2012) 186.

[12] Gao Y., Liang W., Shi Y. and Huang Q., *Phys. A: Stat. Mech. Appl.*, **393** (2014) 579.

[13] Solé R. V., Corominas-Murtra B., Valverde S. and Steels L., *Complexity*, **15** (2010) 20.

[14] Seoane L. F. and Solé R., *Sci. Rep.*, **8** (2018) 10465.

[15] Baechler R. and Seiler G., *Complexity, Isolation, and Variation*, *Linguae & Litterae Series* (De Gruyter) 2016.

[16] Pellegrino F., Marsico E., Chitoran I. and Coupé C., *Approaches to Phonological Complexity*, *Phonology and Phonetics [PP] Series* (De Gruyter) 2009.

[17] Ehret K., *An information-theoretic approach to language complexity: variation in naturalistic corpora*, PhD Thesis, (Albert-Ludwigs-Universität Freiburg) 2016.

[18] Amancio D. R., Aluisio S. M., Oliveira O. N. and da F. Costa L., *EPL*, **100** (2012) 58002.

[19] Koplenig A., Meyer P., Wolfer S. and Müller-Spitzer C., *PLOS ONE*, **12** (2017) 1.

[20] Nettle D., *Philos. Trans. R. Soc. B: Biol. Sci.*, **367** (2012) 1829.

[21] Liu H. and Cong J., *Chin. Sci. Bull.*, **58** (2013) 1139.

[22] Hammarström H., Forkel R., Haspelmath M. and Bank S., *Glottolog 4.3* (Jena) 2020, `https://glottolog.org/accessed 2021-04-08`.

[23] Dryer M. S. and Haspelmath M. (Editors), *WALS Online* (Max Planck Institute for Evolutionary Anthropology, Leipzig) 2013.

[24] Shannon C. E., *Bell Syst. Tech. J.*, **27** (1948) 379.

[25] Bentz C., Alikaniotis D., Cysouw M. and Ferrer-i-Cancho R., *Entropy*, **19** (2017) 275.

[26] Nemenman I., Shafee F. and Bialek W., *Entropy and inference, revisited*, presented at *Advances in Neural Information Processing Systems 14 - Proceedings of the 2001 Conference, NIPS 2001 Advances in Neural Information Processing Systems* (Neural information processing systems foundation) 2002.

[27] Bentz C., Ruzsics T., Koplenig A. and Samardžić T., *A comparison between morphological complexity measures: Typological data vs. language corpora*, in *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)* (The COLING 2016 Organizing Committee, Osaka, Japan) 2016, pp. 142–153.

[28] Brouwer A. and Haemers W., *Spectra of Graphs*, *Universitext Series* (Springer, New York) 2011.

[29] pandas development team, *pandas-dev/pandas: Pandas* (February 2020), `https://doi.org/10.5281/zenodo.3509134`.

[30] McKinney Wes, *Data Structures for Statistical Computing in Python*, in *Proceedings of the 9th Python in Science Conference*, edited by Stéfan van der Walt and Jarrod Millman (SciPy) 2010, pp. 56–61.

[31] Loper E. and Bird S., *Nltk: The natural language toolkit*, in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, Vol. **1**, *ETMTNLP '02* (Association for Computational Linguistics, USA) 2002, p. 63–70, `https://doi.org/10.3115/1118108.1118117`.

[32] Hagberg A. A., Schult D. A. and Swart P. J., *Exploring network structure, dynamics, and function using networkx*, in *Proceedings of the 7th Python in Science Conference*, edited by Varoquaux G., Vaught T. and Millman J. (Pasadena, Cal. USA) 2008, pp. 11–15.

[33] Harris C. R., Millman K. J., van der Walt S. J., Gommers R., Virtanen P., Cournapeau D., Wieser E., Taylor J., Berg S., Smith N. J., Kern R., Picus M., Hoyer S., van Kerkwijk M. H., Brett M., Haldane A., del R'io J. F., Wiebe M., Peterson P., G'erard-Marchant P., Sheppard K., Reddy T., Weckesser W., Abbasi H., Gohlke C. and Oliphant T. E., *Nature*, **585** (2020) 357.

[34] Ferrer-i-Cancho R., Gómez-Rodríguez C., Esteban J. L. and Alemany-Puig L., *Phys. Rev. E*, **105** (2022) 014308.

[35] Bishop C. M., *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer-Verlag, Berlin, Heidelberg) 2006.

[36] Jones E., Oliphant T., Peterson P. *et al.*, *SciPy: Open source scientific tools for Python* (2001), `http://www.scipy.org/`.

[37] van der Maaten L. and Hinton G., *J. Mach. Learning Res.*, **9** (2008) 2579.

[38] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M. and Duchesnay E., *J. Mach. Learning Res.*, **12** (2011) 2825.

[39] Waskom M. L., *J. Open Source Softw.*, **6** (2021) 3021.

[40] Bender E. M., *Linguistically naïve != language independent: Why NLP needs linguistic typology*, in *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?* (Association for Computational Linguistics, Athens, Greece) 2009, pp. 26–32, `https://www.aclweb.org/anthology/W09-0106`.