Laplacian spectrum approach to linguistic complexity: A case study on indigenous languages of the Americas

Article *in* EPL (Europhysics Letters) · March 2020

DOI: 10.1209/0295-5075/129/58003

2 authors:

Javier Vera
Pontificia Universidad Católica de Valparaíso

Wenceslao Palma
Pontificia Universidad Católica de Valparaíso

Laplacian spectrum approach to linguistic complexity: a case study on indigenous languages of the Americas

J. VERA[1] and W. PALMA[1]

[1] *Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile*

PACS 89.20.-a – Interdisciplinary applications of physics
PACS 89.75.-k – Complex systems
PACS 89.75.Fb – Structures and organization in complex systems

Abstract – How to computationally describe the fascinating diversity of indigenous languages of the Americas? This work provides a large-scale and quantitative approach to these languages through the proposal of a novel measure of linguistic complexity based on co-occurrence graphs. Linguistic complexity is measured using the set of eigenvalues obtained from the Laplacian matrix of graphs. The results suggest first that our graph-based definition of linguistic complexity is positively correlated with previous approaches. Second, we were able to describe some structural differences between indigenous languages. We argue that our approach might suggest another application of graph-based techniques to the study of language.

Introduction. – How to computationally describe the fascinating diversity of indigenous languages of the Americas? The answer to this intriguing question involves several issues. First of all, indigenous languages, particularly from the Americas, have received little attention from a computational point of view [1]. This fact is quite surprising because there is an enormous range of language families spoken approximately by 28 million people who self identify as members of an indigenous group in the Americas [2,3]. Moreover, several statistical linguistic universals, establishing deep insights into human language, have been proposed without any consideration of indigenous languages. As an example, the complex morphological paradigm of *Chinantec* language [4, 5] may allow to ask for the validity of the Zipf's law of abbreviation [6].

Secondly, in 2016, the United Nations General Assembly adopted a resolution proclaiming 2019 as the International Year of Indigenous Languages[1]. This resolution was based on that almost half of the languages spoken around the world were in danger of disappearing. Crucially, most of these are indigenous languages and therefore the cultures and knowledge systems to which they belong are put at risk. In this sense, computational studies on indigenous or endangered languages may provide a practical way to increase their web-based visibility.

With the above discussion in mind, our main goal is to define a novel computational approach to measure *linguistic complexity* in order to make global comparisons in a large-scale set of indigenous languages of the Americas. In agreement with [7], irrespective of how we want to measure the complexity of a language, it is essential to keep complexity as an "objective" notion, in the sense of being independent of the use to which we put the system. On this point, we keep complexity apart from user-based notions such as "cost" or "simplification" (for an example of this alternative point of view, see [8]).

According to the *equal complexity hypothesis*, the total complexity of all human languages is considered equal, as proposed by Hockett [9]. Interestingly, as stressed by [10] the equal complexity hypothesis and holistic language typology are complementary through the following question: How to define linguistic parameters which describe a language as a whole? Our work tries to approach the equal complexity hypothesis with great caution and consider this hypothesis as a basic start-point to quantify the comparison between indigenous languages of the Americas. Here, our analyses arise from key questions about linguistic complexity of indigenous languages without any mention of user-centered notions: Are all these languages of equal complexity? Does higher complexity of one module (e.g. *morphology*) imply lower complexity of another module (e.g. *syntax*)? How can complexity be measured? [11].

This work proposes a novel linguistic complexity mea-

[1]https://en.iyil2019.org/

sure based on the Laplacian spectrum of graph-based language representations, in order to make global comparisons between indigenous languages of the Americas. Following the perspective of Comrie [12], who pointed out the need to define parameters which describe language from a holistic and systemic perspective, we describe thus the organization of maps of connections between word items, by using the Laplacian spectrum [13, 14] of co-occurrence graphs representing language data. Within this framework, [15] has revealed, for instance, global properties of the architecture of neural networks, describing the organization of maps of connections between neural elements. The Laplacian spectrum is formed by the set of eigenvalues obtained from the *normalized Laplacian matrix* of a graph. Strikingly, this set of numbers not only capture the global properties of a graph, but also local structures that are produced by graph changes (like motif or node duplication) [16]. Therefore, the Laplacian spectrum allows us to describe the holistic linguistic complexity, particularly across indigenous languages of the Americas.

Recent work on computational approaches to language has proposed remarkable unsupervised information-theoretic measures of linguistic complexity [17, 18]. These measures are mainly based on the average amount of information encoded in word choice using the concepts of entropy and Kolmogorov complexity. The approach developed here is based on the richness of the modeling capabilities of graphs, added up to the wide offer of graph-theoretic or graph-mining algorithms. Furthermore, several works have remarked the fact that language can be modeled by graphs [19–22]. In this paper, linguistic complexity [8, 11, 17, 23] is associated to co-occurrence graphs whose edges capture thus inter-word relationships [24, 25]. To structurally characterize languages, we applied a concept defined over the Laplacian spectrum -the *Laplacian energy*-, as a way to consider a systemic approach to linguistic complexity.

The remaining of the article details the graph-based approach to linguistic complexity of indigenous languages of the Americas. We organize this 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 (indigenous) languages.

Materials and methods. –

Materials. To estimate linguistic complexity, we first need a comparable text corpora across many languages. Ideally, text content must be constant in order to avoid style or genre distortions. To control this constraint across languages, we use a parallel corpus of the *bible.* All the texts were obtained as XML files from [26][2]. 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. From each XML file, we only extracted the text of the *new testament.* Details of the considered indigenous languages are shown in Table 1.

[2]https://github.com/christos-c/bible-corpus

Basic concepts on Graph Theory. We consider an undirected and weighted graph $G = (V, E, W_E)$, completely defined by the vertex set V of size n, the edge set E and the weight set W_E. In our approach, the set V represents word-types for a language, while E is formed by the set of co-occurrences between word-types at distance 1 (that is, bigrams). The *weight* $w(uv) \in W_E$ associated to the edge $uv \in E$ counts the number of co-occurrences of the bigram uv. The *neighborhood* of the node $u \in V$ is the set $V_u = \{v \in V : uv \in E\}$. The (unweighted) *degree* of the node $u \in V$ is simply the size of its neighborhood.

A classical graph-mining tool: clustering coefficient. The notion of *clustering* captures correlations between neighborhoods. In social networks, this notion captures the fact that when there is an edge between two nodes (for example, two individuals are friends) they probably have common neighbors. In mathematical terms, the *average clustering coefficient* for the graph G is defined as:

$$C(G) = \frac{1}{n} \sum_{u \in V} C_u \tag{1}$$

where

$$C_u = \frac{2|vw : v, w \in V_u \ \wedge \ vw \in E|}{|V_u|(|V_u| - 1)}$$

is the *local clustering coefficient* for the node u.

Spectral Graph Theory: an energy function. *Spectral graph theory* is mainly focused on discovering graph properties arising from the eigenvalues of the matrices associated to the graph [28], such as the *adjacency matrix* and the *Laplacian matrix.*

The *normalized Laplacian matrix* is defined by the relation $\mathcal{L} = I - D^{-1/2} A D^{-1/2}$, 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 non-zero vectors u (the corresponding *eigenvectors*) satisfying the equation $\mathcal{L}u = \lambda u$.

The *normalized Laplacian energy* of the graph G is [29]:

$$E_{\mathcal{L}}(G) = \frac{1}{n} \sum_{i=1}^{n} |\lambda_i(\mathcal{L}) - 1| \tag{2}$$

The normalization term $\frac{1}{n}$ allows to compare graphs of different size. As remarked in the introduction, the Laplacian spectrum, and therefore $E_{\mathcal{L}}(G)$, is a simple and efficient way to describe languages at a system level.

Table 1: Basic description of the parallel corpus of indigenous languages of the Americas (based on [26,27]).

ISO 639-3	language	linguistic family	tokens	types	speakers
acu	Achuar	Jivaroan	176499	19240	5000
agr	Aguaruna	Jivaroan	150931	24933	38300
ake	Akawaio	Carib	231633	8631	4500
amu	Amuzgo	Oto-manguean	200928	14185	23000
cjp	Cabecar	Chibchan	201346	8293	8840
cak	Cakchiquel	Mayan	314530	8300	132000
chr	Cherokee	Iroquoian	116733	25537	16400
chq	Chinantec	Oto-manguean	306046	11899	8000
jai	Jakalteco	Mayan	219494	12160	77700
quc	K'iche'	Mayan	272692	7206	1900000
mam	Mam	Mayan	216865	10946	200000
nhg	Nahuatl	Uto-Aztecan	176717	15476	3500
ojb	Ojibwa	Algic	142440	36436	20000
kek	Q'eqchi'	Mayan	256185	8958	400000
quw	Quichua	Quechuan	116883	14944	20000
jiv	Shuar	Jivaroan	138670	23053	46700
usp	Uspanteco	Mayan	226076	8676	3000

Linguistic complexity measure. To define a graph-based measure of linguistic complexity, we apply the notion of *normalized Laplacian energy* $E_{\mathcal{L}}(G)$ to propose the Laplacian-based complexity:

$$C_{Laplacian} = E_{\mathcal{L}}(G) \tag{3}$$

To compare this measure with a classical measure encoding average properties of individual graph elements, we use the average clustering coefficient (from now, *clustering*). Each language is embedded thus into the two-dimensional space $(C_{Laplacian}, clustering)$ in order to find clusters of similar languages.

Graph construction and implementation details. For text preprocessing (whitespace tokenization, punctuation removal and conversion to lower case), we used *NLTK*[3]. Graph-theoretic techniques were made using *NetworkX*[4] and *NumPy*[5]. Source code is available in a public web repository[6].

For each language, the graph G was built along the following steps:

Step 1. Identify the set of bible verses.

Step 2. Preprocess each verse 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 verse in order to find word-type bigrams. Each new bigram between pairs of word-types from W_t defines an edge of the graph. Repetitions of bigrams increase the weight of the respective edge.

Despite of language graphs can be defined in a number of ways (for example, weighted/non-weighted or directed/non-directed), we based our analyses on a simple version of graphs. Further work should discuss the potential influence on the results of graph construction strategies.

Two previous linguistic complexity measures. To compare our graph-based approach with previously defined complexity measures, we introduce two notions. Let T be a text that is drawn from the vocabulary of word-types W_t. Also, we assume that word-type probabilities are distributed according to $p(w)$, $w \in W_t$. The *average information content of word-types* (or the *entropy*) reads [18]

$$C_H = -\sum_{w \in W_t} p(w) \log(p(w)) \tag{4}$$

In addition, *word-type ratio* C_{TTR} is viewed as a simple baseline measure [30]. Higher values of C_{TTR} correspond to higher morphological complexity [18]. This measure is defined as

$$C_{TTR} = \frac{|W_t|}{|T|} \tag{5}$$

where $|W_t|$ and $|T|$ indicate, respectively, the number of word-types and the number of tokens of the text T.

Results. –

Correlations between complexity measures. We apply the non parametric Spearman rank correlation to evaluate relationships between pairs of measures. All correlations reported here are significant at the $p < 0.001$ level.

[3] https://www.nltk.org/
[4] https://networkx.github.io/
[5] https://www.numpy.org/
[6] https://github.com/javiervz/indigenous-languages

First, Figures 1 and 2 display pairwise comparisons between $C_{Laplacian}$, *clustering* and C_H, C_{TTR}. The correlations between $C_{Laplacian}$ and baselines are strongly positive: 0.9 and 0.89 for C_H and C_{TTR}. On the contrary, the correlations between *clustering* and baselines are strongly negative: -0.91 and -0.84 for C_H and C_{TTR}. The results illustrate that our graph-based approach correlates with previous computational approaches to linguistic complexity.

In order to rule out size effects, we describe the correlation between the number of tokens and $C_{Laplacian}$, as shown in Fig. 3. In this case, the correlation is strongly negative, -0.88, while the coefficient of determination is 0.63. It is interesting to notice that for languages with $C_{Laplacian} < 0.1$ there is a large variability in the number of tokens (from 1.5×10^5 to 3×10^5). We found that at least for low complexity languages corpus size effects can be rejected.

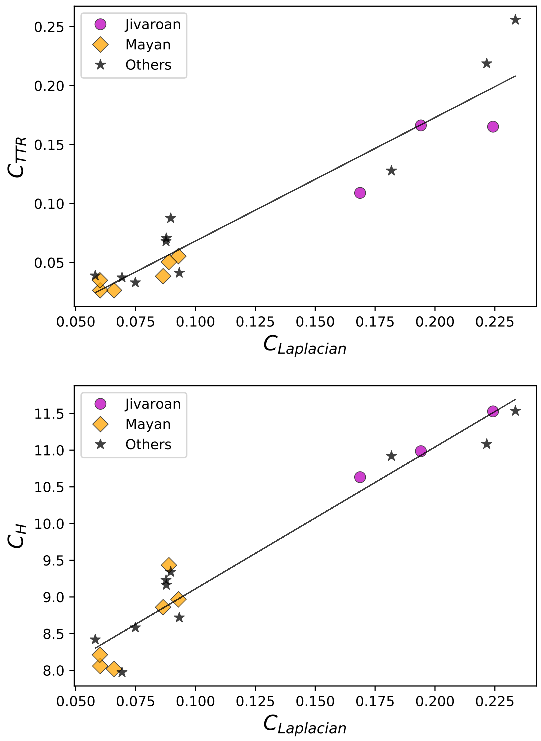

Fig. 1: **Pairwise correlations with $C_{Laplacian}$.** Panels show scatterplots with fitted regression lines. The coefficient of determination is 0.94 and 0.9 for C_H and C_{TTR}. Linguistic families with at least three languages in the corpus are represented by specific symbols.

Indigenous languages in the two-dimensional space. To quantitatively describe indigenous languages of the Americas, from each graph-based representation we define Laplacian-based and clustering axes. Languages were embedded thus into the two-dimensional space, as shown in Fig. 4. Strikingly, this figure illustrated two important issues. First, language families were clustered by linguis-

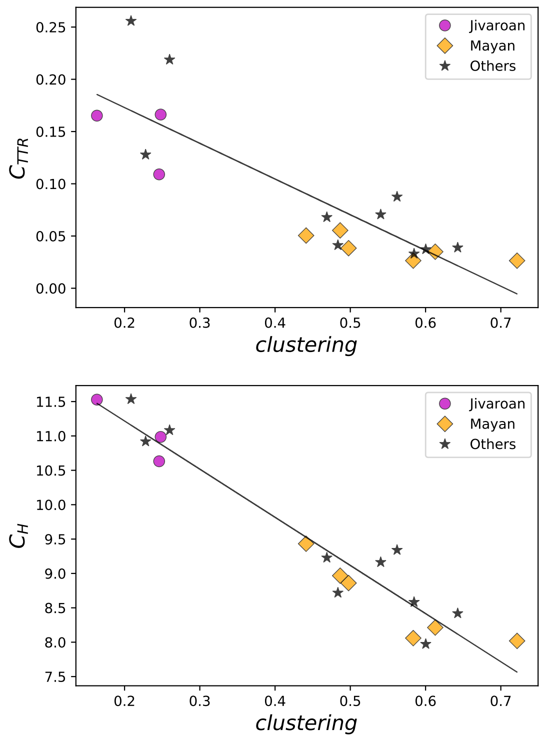

Fig. 2: **Pairwise correlations with *clustering*.** Panels show scatterplots with fitted regression lines. The coefficient of determination is 0.92 and 0.72 for C_H and C_{TTR}.

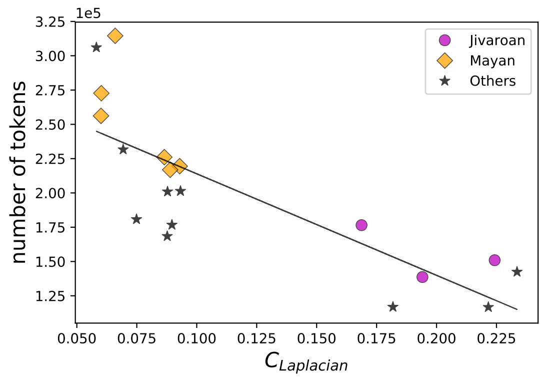

Fig. 3: **Correlation between the number of tokens and $C_{Laplacian}$.** Panel shows a scatterplot with the fitted regression line. The coefficient of determination is 0.63.

tic complexity values: *Jivaroan* and *Mayan* languages exhibited obvious differences in the complexity space. This result is closely related to the work of [31], which proposed a method that first builds a co-occurrence representation of parallel corpora and then extracts graph-mining measures as features for unsupervised machine learning. Interestingly, the Laplacian spectrum (possibly formed by thousand of numbers) can be used as a vector of features within the mentioned approach.

Our evidence suggested also the appearance of a linear negative relationship between $C_{Laplacian}$ and *clustering*.

This fact supported positive evidence for the *equal complexity hypothesis* [9]. In this sense, it is interesting to remark opposite languages: on the one hand, *Chinantec*; and, on the other hand, *Ojibwa*, *Aguaruna* or *Cherokee*.

To empirically test the validity of the *equal complexity hypothesis*, we proposed a simple overall complexity function defined as the sum of $C_{Laplacian}$ and *clustering*. Values are normalized by the maximum: if the hypothesis is true, we expected to find an overall value close to 1. Figure 5 displays the overall complexity for indigenous languages of the Americas (including *English* and *Spanish*).

We contrast the results with the "null hypothesis" that $C_{Laplacian}$ and *clustering* are always negatively correlated whatever the underlying graph is. To do this, we compare the original results with: (a) randomized versions of texts; (b) unweighted random graphs; and (c) weighted random graphs ((b) and (c) defined by expected degree sequence [32]). First, black line displays $C_{Laplacian} + clustering$ for randomized versions of texts. Interestingly, the overall complexity linguistic function behaves as in the original texts. We argue that these results were resilient to text randomization because co-occurrence counts are strongly related to word-frequency distributions. This may explain the positive correlation between Laplacian-based complexity and language entropy [18], as shown in Fig. 3. Word choice distribution can be described thus by information-theoretic and graph-mining techniques. Further work should be conducted to establish relations between these two points of view about language. Random graphs display interesting facts. First, across languages the average value of the overall complexity function is: 1.13 ± 0.09 (original); 1.21 ± 0.07 (a); 1.25 ± 0.08 (b); and 1.43 ± 0.14 (c). Second, random graphs are less resilient than random texts. Third, it seems that unweighted graphs lose the inverse relationship between $C_{Laplacian}$ and *clustering*. This suggests that graph structure is embedded in co-occurrence counts (that is, graph weights).

Clustering coefficient and linguistic complexity. Within our approach to linguistic complexity, languages can be conceptualized as a network of words interrelated with each other in complex ways. Do they all relate each other, or are they maybe separated from each other? A sketch of answer is provided by a closer analysis of the linguistic neighborhoods of words. For this, we focused on the local clustering coefficient C_u defined in Eq. 1, measuring the local density of edges in word u's neighborhood.

To quantitative describe the behavior of local word's neighborhoods for each language, we first identify the set of different values of node degree. Then, for each value d we calculated $\frac{1}{|V_d|}\sum_{u \in V_d} C_u$, where V_u is the set of nodes with degree d. As shown in Fig. 6, we first observed that the three considered languages, exhibiting respectively high, middle and low values of $C_{Laplacian}$, showed a different behavior for *clustering* versus node degree. Indeed, for *Kiche* the local clustering coefficient is very large in comparison with the other languages. For example, for words with 100 neighbors, the average local clustering is 0.25, 0.11 and 0.07 respectively for *Kiche*, *Achuar* and *Aguaruna*. Meanwhile, our analysis also showed that *clustering* decreases monotonically with node degree. Particularly, the clustering coefficient drops linearly (in $\log - \log$ scale) for words with more than 10^3 neighbors, indicating that these words are likely forming part of the language graph for several non-necessarily related linguistic mechanisms. Interestingly, a similar behavior has been previously observed in social networks [33].

Discussion. –

Graph-based complexity across indigenous languages of the Americas. In this short paper, we described a novel computational measure of linguistic complexity, applied in a large-scale collection of indigenous languages of the Americas. Our approach was mainly based on the broader interest to develop computational studies and Natural Language Processing tools on these languages. In this sense, we agree with [1] that technological and computational approaches can have a positive social impact for the communities which depend on these languages; and in addition the great diversity of indigenous languages of the Americas posses fascinating scientific challenges.

Co-occurrence graphs and syntactic dependencies. Here, we developed a co-occurrence approach to capture complexity issues of language. A possible criticism of this approach is that it is mainly based on word-frequency matters: graph edges are defined by repetitions of consecutive pairs of words. We may ask therefore if this local strategy of capture short-range word relations fulfills the goal of compare languages at a global level. To shed light on this problem, we remark that several works have showed (for example, [34–36]) that (1) average distance between words is small; and (2) it is a very slowly growing function of sentence length. The euclidean distance is defined between syntactically linked words in sentences. We argue that these two findings allow us to build co-occurrence graphs based only on local information about pairs of words. Despite that some pairs can be spurious, part of the structure of language is captured by co-occurrence graphs.

Indigenous languages into the complexity space. In Subsection "Indigenous languages in the two-dimensional space", we embedded indigenous languages of the Americas into the two-dimensional space formed by $C_{Laplacian}$ and *clustering*. We found that linguistic families, for example *Mayan* [37] and *Jivaroan* [38] languages, are located in separated clusters. This is to be expected, since both graph-based measures are strongly correlated with previous approaches, although the appearance of only two clearly separated clusters requires further typological and linguistic work. One fact is more surprising. Results suggested that it is a negative linear trend of languages into the complexity space. In theory, it may be natural to think in a (linear) relationship between the clustering coefficient and the Laplacian energy of graphs. However, the overall

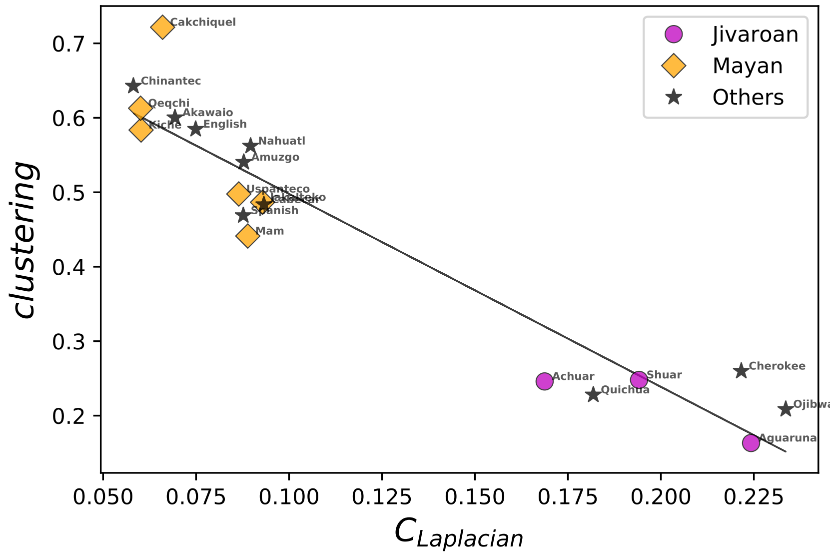

Fig. 4: **Complexity space for indigenous languages of the Americas.** Axes indicate values for each graph-based measure: $C_{Laplacian}$ and *clustering*. Black line represents a fitted regression line with coefficient of determination $r^2 = 0.9$.

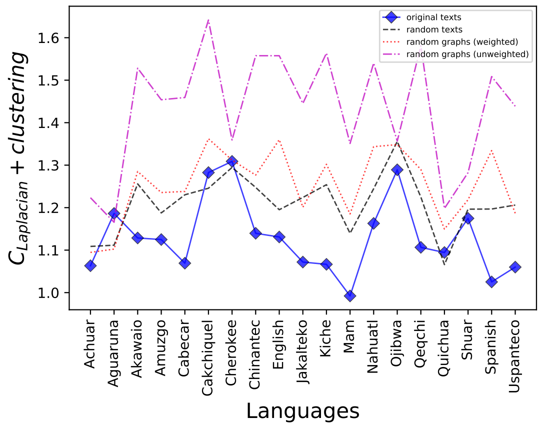

Fig. 5: **Overall complexity for indigenous languages of the Americas.** Languages are associated to a simple measure of overall complexity: $C_{Laplacian} + clustering$. Blue line indicates original values. Black line represents the overall complexity for random versions of each text. Random graphs are represented by two lines: orange (weighted) and purple (unweighted). For each measure, values are normalized by the maximum.

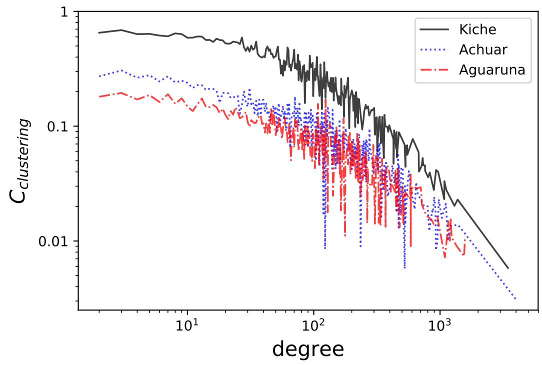

Fig. 6: **Local clustering coefficient as a function of degree.** For languages with high, middle and low values of $C_{Laplacian}$, *Kiche*, *Achuar* and *Aguaruna*, the panel shows the average value of *clustering* for each node degree.

complexity (simply defined as the sum of $C_{Laplacian}$ and *clustering*) remains approximately constant (and close to 1) across languages. This suggests a positive evidence for the *equal-complexity hypothesis*, establishing that despite their evident structural differences languages exhibit an equal overall complexity level [10].

Linguistic features and graph-based representations. The holistic comparison between languages was reflected by embedding languages into a complexity space, whose axes are $C_{Laplacian}$ and *clustering*. These axes are defined using graph-theoretical tools that do not take into consideration microscopic linguistic features (using, for example, [39]). Several interrelated ideas arise from this linguistically-agnostic approach. We stress indeed the fact that the application of graph-mining techniques on holistic representations of languages (particularly, using graphs) opens new research questions about the "meaning" of those techniques. For example, we may ask for the linguistic interpretation of a large or small average clustering coefficient: How this behavior affects microscopic linguistic features? What happens in indigenous languages? Moreover, we should question about which graph-mining techniques can be used as representations of linguistic complexity. The answer is not obvious. In practical terms, there are at least two possible ways: (1) the spectral graph-theoretic point of view extracts several (maybe thousands) of eigenvalues to describe a graph-based representation of language. Linguistic complexity is therefore a bag of numbers (or a vector) or a number extracted from this bag (for example, our energy-based approach); the other way (2) follows the path established by many popular graph-mining metrics, that encode average properties of individual graph elements. An interesting work in this line is [31], in which co-occurrence representation of languages allowed the extraction of several graph-mining measures in order to apply machine learning models. We argue that linguistic complexity should be analyzed by complementing approaches (1) and (2).

REFERENCES

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

[2] Sebeok T., *Native Languages of the Americas* no. v. 1 (Springer US) 2013.

[3] Wagner C., *Documentos Lingüísticos y Literarios*, **17** (1991) 30.

[4] Palancar E. L., *Revisiting the complexity of the Chinantecan verb conjugation classes* in *Patterns in Meso-American Morphology* 2014 pp. 77 – 102.

[5] Baerman M. and Palancar E., *The organization of chinantec tone paradigms* in *Proccedings of the Décembrettes. 8th International conference on morphology. December 6-7, 2012*, edited by Augendre S., Couasnon-Torlois G., Lebon D., Michard C., Boyé G. and Montermini F., Carnets de grammaire 2014 pp. 46 – 59.

[6] Ferrer-i-Cancho R., *Complexity*, **21** (2016) 409.

[7] Dahl Ö., *The Growth and Maintenance of Linguistic Complexity* Studies in language Amsterdam / Companion series (John Benjamins) 2004.

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

[9] Hockett C. and F H., *A Course in Modern Linguistics* A Course in Modern Linguistics (Macmillan) 1958.

[10] Oh Y. M., *Linguistic complexity and information : quantitative approaches* Ph.D. thesis université Lyon 2 (2015).

[11] Baechler R. and Seiler G., *Complexity, Isolation, and Variation* linguae & litterae (De Gruyter) 2016.

[12] Comrie B., *Language Universals and Linguistic Typology: Syntax and Morphology* (University of Chicago Press) 1989.

[13] Banerjee A. and Jost J., *Linear Algebra and its Applications*, **428** (2008) 3015 .

[14] Banerjee A. and Jost J., *Discrete Applied Mathematics*, **157** (2009) 2425 networks in Computational Biology.

[15] de Lange S., de Reus M. and Van Den Heuvel M., *Frontiers in Computational Neuroscience*, **7** (2014) 189.

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

[17] Ehret K., *An information-theoretic approach to language complexity: variation in naturalistic corpora* Ph.D. thesis albert-Ludwigs-Universität Freiburg (2016).

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

[19] Cong J. and Liu H., *Physics of Life Reviews*, **11** (2014) 598 .

[20] Gao Y., Liang W., Shi Y. and Huang Q., *Physica A: Statistical Mechanics and its Applications*, **393** (2014) 579 .

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

[22] Seoane L. F. and Solé R., *Scientific Reports*, **8** (2018) 10465.

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

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

[25] Nettle D., *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367** (2012) 1829.

[26] Christodouloupoulos C. and Steedman M., *Language Resources and Evaluation*, **49** (2015) 375.

[27] Eberhard D. M., Simons G. F. and Fennig C. D., (Editors) *Ethnologue: Languages of the World* twenty-second Edition (SIL International, Dallas, TX, USA) 2019.

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

[29] Cavers M., Fallat S. and Kirkland S., *Linear Algebra and its Applications*, **433** (2010) 172 .

[30] Kettunen K., *Journal of Quantitative Linguistics*, **21** (2014) 223.

[31] Liu H. and Cong J., *Chinese Science Bulletin*, **58** (2013) 1139.

[32] Chung F. and Lu L., *Proceedings of the National Academy of Sciences*, **99** (2002) 15879.

[33] Ugander J., Karrer B., Backstrom L. and Marlow C., *The anatomy of the facebook social graph* (2011).

[34] Ferrer-i-Cancho R. and Solé R. V., *Proceedings. Biological sciences*, **268** (2001) 2261 11674874[pmid].

[35] Ferrer-i-Cancho R., *Phys. Rev. E*, **70** (2004) 056135.

[36] Ramon F. and Haitao L., *glot* Vol. 5 2014 Ch. The risks of mixing dependency lengths from sequences of different length p. 143 2.

[37] Aissen J., England N. and Maldonado R., *The Mayan Languages* Routledge Language Family Series (Taylor & Francis) 2017.

[38] Adelaar W. and Muysken P., *The Languages of the Andes* Cambridge Language Surveys (Cambridge University Press) 2004.

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