LETTER

The community structure of word co-occurrence networks: Experiments with languages from the Americas

To cite this article: Javier Vera and Wenceslao Palma 2021 *EPL* **134** 58002

A Letters Journal Exploring the Frontiers of Physics

June 2021

EPL, **134** (2021) 58002 www.epljournal.org
doi: 10.1209/0295-5075/134/58002

The community structure of word co-occurrence networks: Experiments with languages from the Americas

Javier Vera[(a)] and Wenceslao Palma

Pontificia Universidad Católica de Valparaíso - Valparaíso, Chile

received 5 May 2021; accepted in final form 8 June 2021
published online 10 August 2021

Abstract – We study a set of algorithms to discover the community structure of networks for languages from the Americas. Our experiments are based on a parallel corpus which allows us to represent each language as a co-occurrence network. Four methods to calculate network modularity, as a measure of the quality of community structure, were used. We studied several aspects of the community structure of co-occurrence networks. First, we were able to construct the map of modularity variations across languages from the Americas. With this, we separated large groups of languages into low- and high-modularity families. We suggested also a strong influence of functional words on low-modularity languages. Finally, we found a strong relationship between word entropy values and modularity. Our approach is thus a simple network-based contribution to face data scarcity of languages which are in danger of disappearing.

Introduction. – Is it possible to describe the formation of densely connected groups of words, which are sparsely connected between them? Is there any linguistic interpretation of such groups? What happens in languages from the Americas? This paper is an attempt to answer these questions. To do this, we deal with a network-based account for languages from the Americas. Remarkably, most of these languages are less privileged, endangered and resource scarce (*low-resource* languages in terms of [1]). In this sense, our work proposes a network-based approach to languages which traditionally have not received much attention from a Statistical Mechanics or a computational point of view.

With the widespread of Natural Language Processing (NLP), Computational Linguistics and Statistical Mechanics approaches to language problems, it is important to take fairness issues into consideration while studying human communication systems (see [2,3] for recent surveys on fairness and political implications of machine learning). A recent paper [4] has identified, for example, evident gender biases in word embeddings. Working with languages from the Americas would help to language fairness, understood here as the absence of favoritism towards some languages (like *English*) based on their inherent characteristics.

Following [5], a property that is common to many networks is *community structure*, defined as the division of network nodes into densely connected groups, which are sparsely connected between them. Searching for such groups can provide invaluable help in understanding and visualizing the structure of networks. It is natural to ask thus for the meaning of community structure in the context of co-occurrence networks whose nodes represent linguistic units (like words).

Recently, human language has been studied within a Statistical Mechanics approach (see, for example, [6–9]). In this line, two recent intriguing works have contributed to the study of large-scale text analysis. Perc [10] noticed that the appearance of *Zipfian* properties in a dataset can be understood as an indication of large-scale self-organization patterns. In this sense, this paper provided evidence in order to stress the fact that preferential attachment processes, like the Matthew effect [11], played a central role in determining the emergence of scaling in text corpora. In their article [12], they studied the structure and complexity of texts in *Slovene belles-lettres*. Using a co-occurrence network approach, they found clear distinctions in the statistical properties between different age groups.

Remarkably, network-based models have been used in several language applications. Using Multilayer networks [13], have addressed the problem of multi-document extractive summarization for both *English* and *Portuguese* languages. To face the problem of authorship attribution [14], applied the concept of motifs, understood as recurrent interconnection patterns. As mentioned by the

[(a)]E-mail: javier.vera@pucv.cl (corresponding author)

authors of this work, their findings proposed a novel methodology to be further explored in other tasks (as language characterization). Here, languages are viewed as co-occurrence networks whose edges capture thus inter-word relationships [15,16]. In line with [17], in which machine learning models are based on the application of network-mining tools on word co-occurrence networks, we aim, on the one hand, at quantifying the strength of the community structure across languages from the Americas; and, on the other hand, to provide an interpretation of this strength in terms of a previous account based on information-theoretic entropy of natural languages [18].

The remaining of the article details our network-based approach to quantifying the strength of community structure across languages from the Americas. We organize the discussion in three sections. The next section, "Materials and methods", describes language data and the technical notions of community structure. Section "Results" describes and illustrates the main results. Section "Discussion" summarizes our work and restates the key challenges of our approach to the community structure in low-resource languages.

Materials and methods. –

Universal declaration of human rights corpus UDHR. To avoid style or genre distortions across languages, we based our experiments on the parallel corpus of the *universal declaration of human rights*[1] (*UDHR*). We need some basic notation. A *word type* is a unique string delimited by white spaces (*the*, *cat*, *chases* and *a* in *the cat chases a cat*). A *word token* is then any repetition of a word type. Details about the *UDHR* corpus are shown in fig. 1 and table 1.

Information-theoretic entropy. As suggested by the seminal works of Shannon [19], 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 [19]

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

Using several corpora and tackling some problems of word entropy estimation [18], provided a public database of entropy values for 1259 languages. Since all entropy estimators are strongly correlated, for our experiments we used the entropy values provided by the *NSB* estimator [20].

For the large sample of languages of the world [18], 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 [22].

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

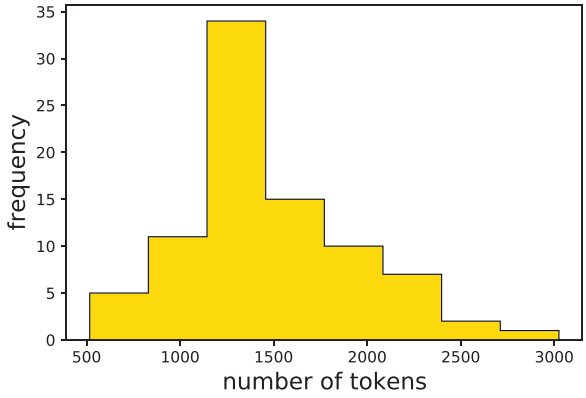

Fig. 1: Basic description of the parallel corpus of languages of the Americas. The figure displays the histogram of the number of tokens for each text of the *UDHR* corpus.

Table 1: Basic description of the linguistic families from the Americas (based on *Glottolog* data [21]). Languages analyzed here appear in both the *UDHR* corpus and *Glottolog*.

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

Basic concepts on Network Theory. We consider an undirected network $G = (V, E)$, where V represents the set of nodes of size n and E is the set of edges. In our approach, the set V represents word types for a language, while E is formed by pairs of adjacent word types. 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. The *weight* $w(uv) \in W_E$ associated to the edge $uv \in E$ counts the number of appearances of the bigram uv.

A key graph-theoretic measure to detect clusters of related nodes (word types or categories) is *modularity* [5]. This measure arises from the fact that a good community structure for a graph corresponds to a statistically surprising arrangement of edges. More formally, the modularity Q of a graph $G = (V, E)$ is a scalar number between -1 and 1 that measures the density of links inside

communities as compared to links between communities. This measure is defined as [23]

$$Q = \frac{1}{2m} \sum_{ij \in E} [a_{ij} - P_{ij}] \delta(c_i, c_j), \tag{2}$$

where $P_{ij} = k_i k_j / 2m$ is the expected number of edges between i and j for a random null model. $\delta(c_i, c_j) = 1$ if the nodes i and j belong to the same cluster, and 0 otherwise.

Finding community structure. Following [5], the community structure of a network G is the division of network nodes into densely connected groups, sparsely connected between them. More precisely, the *community structure* of G is a mapping between the node set V and a set of *Labels* that assigns to each node one label in *Labels*. Our work is based on four ways of finding such mapping.

i) *Louvain algorithm.*

The *Louvain algorithm* [24] quickly finds high modularity partitions unfolding a complete hierarchical community structure for the graph. The algorithm is divided into two phases: a first phase of modularity maximization (2), followed by a phase in which nodes are aggregated into communities.

ii) *Leiden algorithm.*

The *Leiden algorithm* [25] is a recently proposed improved version of the *Louvain algorithm*. The key aspect of this algorithm is that it guarantees the connection between the obtained communities. Each step of the *Leiden* algorithm consists in the local moving of nodes, the refinement of the partition and the aggregation of the network based on the refined partition.

iii) *Eigenvector.*

This algorithm is based on the spectral properties of the so-called modularity matrix [26].

iv) *Surprise communities.*

This method [27] is based on the development of an efficient algorithm for optimizing surprise [28].

Network construction and implementation details. For basic NLP preprocessing (whitespace tokenization, punctuation removal and conversion to lower case), we used *NLTK* [29]. Network-theoretic techniques were made using *NetworkX* [30] and *CDlib* [31].

Each language is represented by a co-occurrence network, and then by 4 values of modularity. The network G was built along the following steps:

Step 1. Preprocess each sentence by whitespace tokenization, punctuation removal and conversion to lower case.

Step 2. Define the set of word types W_t of the entire text.

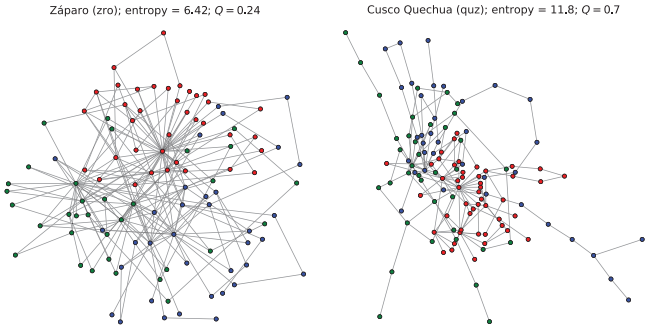

Fig. 2: Co-occurrence networks for two languages of the Americas. The figure displays co-occurrence networks for languages with low and high *Leiden* modularity: *Záparo* (zro) and *Cusco Quechua* (quz). To compare with previous accounts, word entropy values are presented (taken from [18]). For each network, only the three largest communities are showed. Colors indicate community labels.

Step 3. Through an iterative process, inspect each sentence in order to find word-type adjacent co-occurrences (based on the fact that dependency relationships occur in general at small distances [32]). Each new co-occurrence between pairs of word types from W_t defines an edge of the network. Repetitions of bigrams increase the weight of the respective edge.

Results. –

Measuring community structure across languages from the Americas. We first shed light on network-based visualizations of some languages from the Americas. Figure 2 displays co-occurrence networks for languages with low and high *Leiden* modularity: *Záparo* (zro) and *Cusco Quechua* (quz). As expected, there are radical structural differences between languages associated to such modularity values. A strong community structure is observed for *Cusco Quechua*; by contrast, the co-occurrence network for *Záparo* exhibited little evidence of community structure. As shown in the figure, word entropy values are positively correlated with modularity. Remarkably, previous studies (for example, [22]) have stressed the fact that word entropy is positively correlated with morphological complexity. With this, it seems reasonable to think that morphological complexity is related to a strong community structure.

Figure 3 displays density plots for the estimated modularities for all 85 languages from the Americas. Modularity is distributed as follows: (*Louvain*) around a mean of 0.51 (SD = 0.11); (*Leiden*) around a mean of 0.45 (SD = 0.14); (*Eigenvector*) around a mean of 0.38 (SD = 0.13); and (*Surprise*) around a mean of 0.36 (SD = 0.14). It is clear that the modularity obtained by the *Louvain* algorithm is slightly higher than the other network-mining techniques.

To visually illustrate the areality of modularity across languages from the Americas, fig. 4 provides a map with modularity values obtained by the *Louvain* algorithm. The range of values is indicated by a three-way scale of

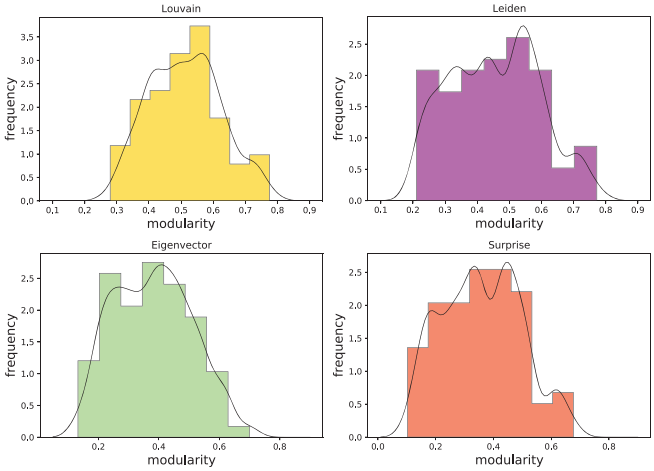

Fig. 3: Modularity across languages from the Americas. The figure displays modularity histograms across languages from the Americas. Languages are represented by co-occurrence networks. With this, modularities are obtained using different network-mining techniques: *Louvain Leiden*, *Eigenvector* and *Surprise*. Black lines indicate a Kernel-density estimation [33] with bandwidth of 0.35.

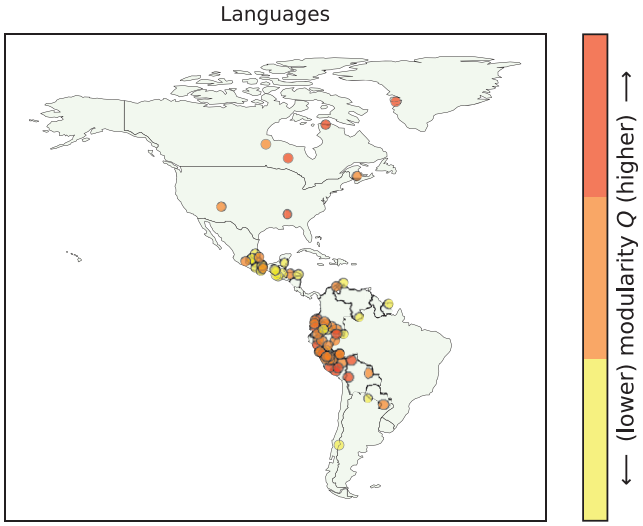

Fig. 4: Map with modularities across languages from the Americas. The figure displays *Louvain* modularity values of fig. 3 across languages from the Americas. Our sample of languages is focused on *Mesoamerican* and *North Andean* regions. The numeric range is divided into three parts: high, middle and low. Darker color indicates higher values of modularity.

high, medium and low modularities. As can be seen in the map, there are high- and low-modularity areas across the Americas, coinciding respectively with high and low word entropy areas as previously remarked by [18]. For example, languages in the Andean region of South America exhibit high modularities. By contrast, languages from Mesoamerica (like the *Otomanguean* family) exhibit relatively low modularities. To quantitatively confirm part of these observations, we calculated the mean of modularity for some large linguistic families from the Americas (using the *Louvain* algorithm): *Panoan* 0.52 (SD = 0.07); *Arawakan* 0.52 (SD = 0.03); *Mayan* 0.39 (SD = 0.02); *Otomanguean* 0.4 (SD = 0.0); and *Quechuan* 0.62 (SD = 0.05).

Communities and the influence of functional words. There are two intriguing questions arising from the above results: what is the meaning of communities across languages of the Americas? Is there any influence of functional words on the community structure? To propose an answer to these questions, we need to consider the empirical Zipf's law [34,35], which establishes a dichotomy between low-memory words (like the word "the") and low-ambiguity words (like the word "cat"). At least for *English*, within a statistical point of view, text corpora evidence strong scaling properties in word frequencies (see for example [10,11,36,37]). The existence of scaling properties is directly related to the removal of "functional" words (the so-called *stopwords*) in computational tasks. As an example, in [38] keyword extraction is based on co-occurrence modeling of (*English*) texts. As usual, they performed several preprocessing steps, including stopword removal. Is this possible in languages from the Americas? Are there scaling properties across languages from the Americas?

We performed a simple experiment: we removed the top 5% most frequent word types of each text. The underling hypothesis is that modularity values of co-occurrence networks for morphologically complex languages (like the *Quechuan* family) are more robust to stopword removal than morphologically simpler languages (like the *Zaparoan* family). We used modularity values obtained by the *Leiden* algorithm (denoted as Q). To test the hypothesis, we measured *Leiden* modularity after stopword removal. We denoted as Q^*/Q the ratio between the modularities obtained from restricted and original texts. Increasing values of Q^*/Q denote more drastic effects of stopword removal.

As shown in fig. 5, there is a clear exponential decay of Q^*/Q as Q increases. This fact suggested in principle that the community structure for high-modularity families (like *Quechuan*) is more robust under stopword removal. Put differently, for such high-modularity families it is not possible to propose a clear definition of stopwords. The crucial aspect of this observation is that the meaning of the communities changes over modularity values. It is reasonable to hypothesize thus that the detection of lower modularity (as exhibited by *Mayan*) is a strong evidence of function words, which are shared by several communities. This would explain why Q^*/Q increases as Q decreases. Future work should confirm these observations in other languages of the world.

Low-dimensional representations of modularity. To describe the distribution of languages using modularity values, we applied the *t-SNE* [39] dimensionality reduction technique (using the *scikit-learn* implementation [40]). In our case, we reduced the dimensionality of the four

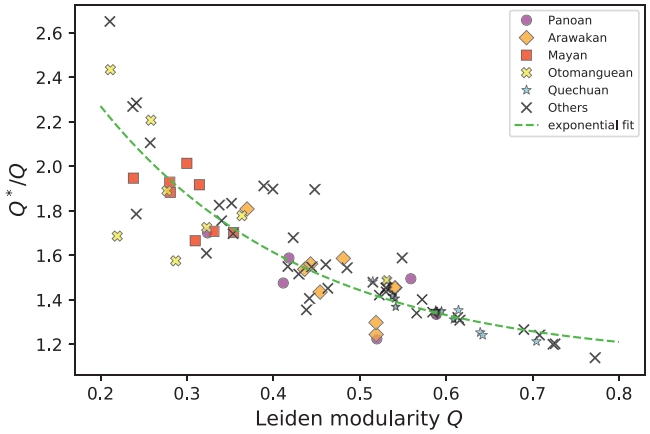

Fig. 5: Influence of the removal of the top 5% most frequent words on modularity. The figure displays *Leiden* modularity Q *vs.* the ratio between Q^* (modularity after stopword removal) and Q. To define stopwords, we followed a simple rule: from the original texts, we removed the top 5% most frequent words. For each language, restricted versions of their co-occurrence networks were constructed. Green depicted line indicates the exponential fit $Q^*/Q \sim 2.7 \exp^{-4.2Q} + 1.10$.

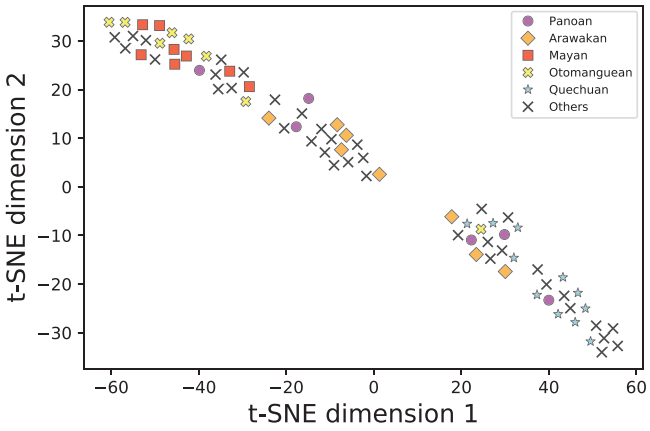

Fig. 6: Two-dimensional representation of modularities using *t-SNE*. Languages are represented by the modularity values obtained by four methods: *Louvain*, *Leiden*, *Eigenvector* and *Surprise*. With this, according to our corpus of 85 languages from the Americas' largest linguistic families (*Panoan*, *Arawakan*, *Mayan*, *Otomanguean* and *Quechuan* languages) are plotted in the two-dimensional space formed by *t-SNE* embeddings. Other languages are indicated with black ×.

modularity values to two. As shown in fig. 6, some previous observations can be confirmed. It is clear that low-modularity families (like *Mayan* or *Otomanguean*) are grouped together and separated from high-modularity families (like the *Quechuan* family).

To provide a quantitative approach to the relationship between word entropy and modularity, we trained a multivariate *Random Forest* (RF) regression [40] that takes the four modularities representing each language in order to predict word entropy. The RF algorithm is one of the best algorithms for classification. The basic concept

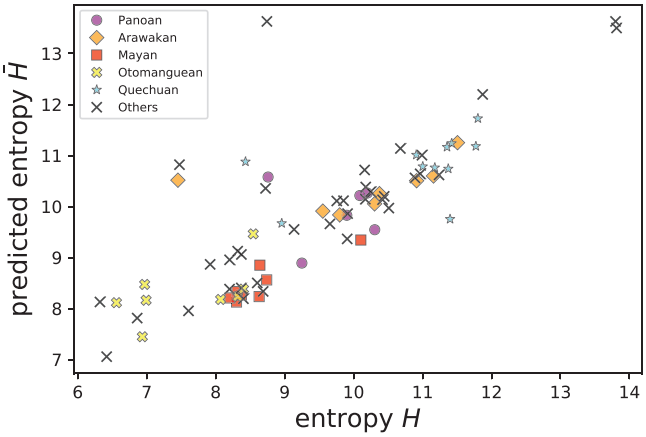

Fig. 7: Entropy H *vs.* predicted entropy $\bar{H}$. The figure displays word entropies (taken from [18]) *vs.* predicted entropies obtained by a multivariate RF regression for modularity values (obtained by four methods). More precisely, we trained a RF that accepts modularity values (for each language, a vector of dimension 4) and predicts word entropy. This suggests a strong relationship between word entropy and modularity values.

is that a group of "weak learners" may come together to build a "strong learner" [41]. We fitted 100 decision trees (train set of size 0.25). To measure our predictions, we used the *Mean Absolute Percentage Error*, $MAPE = \frac{100}{85} \sum_{85} \frac{|H(L) - \bar{H}(L)|}{H(L)}$, where 85 is the size of our sample of languages from the Americas, $H(L)$ represents the word entropy obtained by [18], and $\bar{H}(L)$ is the RF prediction based on modularities of the language L. With this, the *accuracy* of our prediction is the *MAPE* subtracted from 100%. In summary, our model has learned to predict word entropy values with 82.1% *accuracy*. This fact establishes an interesting relationship between morphologically complex languages (evidenced by higher values of word entropy) and a strong community structure.

Discussion. –

Community structure of co-occurrence networks of languages from the Americas. In this work, we described the community structure of co-occurrence networks representing languages from the Americas. Our network-based approach attempted to extract typologically valuable information from parallel corpora. To do this, we ran experiments in which modularities were calculated using four different methods. As remarked in previous sections, we concluded two key ideas. First, there is a clear positive correlation between word entropy and modularity. Moreover, we can predict word entropy only based on modularities. This suggested that languages associated to evident morphological complexity (like the *Quechuan* family) exhibit at the same time a clear community structure for their word organization (expressed here through co-occurrence networks). In this sense, it is interesting to notice also the appearance of an exponential decay of

the ratio between Q^*/Q *vs.* Q, as a key evidence of the robustness of high-modularity families against stopword removal.

Modularity diversity across languages from the Americas. We estimated modularities for a sample of 85 languages from the Americas. Particularly, we found that (*Louvain*) modularities are around 0.5. It is surprising moreover that languages from the Americas fall in a wide spectrum of modularities: from a weak evidence of community structure (as observed in some *Mayan* languages) to large modularities (like *Quechuan* languages). According to [23], if the number of within-community edges is no better than random, the modularity is close to 0; by contrast, values approaching 1 indicate strong community structure. Modularity values (depending on the considered method) range approximately from 0.2 to 0.8. The distributions are also skewed to the right, suggesting that languages of large modularity are scarce. Further work should clarify this fact studying languages from other areas of the world.

NLP systems and the promise of Big Data. A brief chronicle of a failure foretold. In order to confirm previous observations about the relationship between morphological complexity (measured by word entropy [18] diversity across languages of the world) and modularity, we attempted to ask to what extent modularity is related to specific features of the World Atlas of Language Structures (WALS) [42]. We choose 2 chapters/features of WALS which are relevant for describing morphology ([22] used 28 chapters/features of WALS): 20A [43] (Fusion of Selected Inflectional Formatives) and 22A [44] (Inflectional Synthesis of the Verb). The first feature is related to a universal scale of less to more tightly packed word forms, observed respectively in the opposition between isolating and agglutinative/polysynthetic languages. The second feature relies on the degree of synthesis that can be used in the verbs, measuring the category-per-word rate. With this, our aim was to compare a corpus-based approach (modularity measures) with a typology-based approach to morphological complexity. However, WALS only provided information about 2 of the languages of our corpus. In consequence, this does not seem to be feasible on the basis of existing WALS data.

What can we learn from this failure? Several interesting ideas must be remarked. In the first place, languages from the Americas are low-resource languages (LRLs) [1]. Crucially, most of the LRLs are indigenous languages in which the cultures and knowledge systems to which they belong are put at risk. Secondly, it is important to remark that despite one can speak about *Big Data* hidden in data scarcity there are cultures, knowledge systems and languages in danger of disappearing. Some languages (like *English*) have the *privilege* of been studied from a computational point of view. Put differently, these languages have for example the privilege of being annotated in WALS.

Data availability statement: The data that support the findings of this study are openly available at the following URL/DOI: `https://www.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 preprint, arXiv:2006.07264 (2020).
[2] Mehrabi N., Morstatter F., Saxena N., Lerman K. and Galstyan A., *ACM Comput. Surv. (CSUR)*, **54** (2021) 1.
[3] Crawford K., *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (Yale University Press) 2021, `https://books.google.cl/books?id=XvEdEAAAQBAJ`.
[4] Bolukbasi T., Chang K.-W., Zou J., Saligrama V. and Kalai A., *Man is to computer programmer as woman is to homemaker? debiasing word embeddings*, in *Proceedings of the 30th International Conference on Neural Information Processing Systems* NIPS'16 (Curran Associates Inc., Red Hook, NY, USA) 2016, pp. 4356–4364.
[5] Newman M. E., *Proc. Natl. Acad. Sci. U.S.A.*, **103** (2006) 8577.
[6] Cong J. and Liu H., *Phys. Life Rev.*, **11** (2014) 598.
[7] Gao Y., Liang W., Shi Y. and Huang Q., *Phys. A: Stat. Mech. Appl.*, **393** (2014) 579.
[8] Solé R. V., Corominas-Murtra B., Valverde S. and Steels L., *Complexity*, **15** (2010) 20.
[9] Seoane L. F. and Solé R., *Sci. Rep.*, **8** (2018) 10465.
[10] Perc M., *J. R. Soc. Interface*, **9** (2012) 3323.
[11] Perc M., *J. R. Soc. Interface*, **11** (2014) 20140378.
[12] Markovič R., Gosak M., Perc M., Marhl M. and Grubelnik V., *J. Complex Netw.*, **7** (2018) 114.
[13] Tohalino J. V. and Amancio D. R., *Phys. A: Stat. Mech. Appl.*, **503** (2018) 526.
[14] Marinho V., Hirst G. and Amancio D., *Authorship attribution via network motifs identification*, in *Proceedings of the 2016 5th Brazilian Conference on Intelligent Systems (BRACIS)* (IEEE Computer Society, Los Alamitos, Cal., USA) 2016, pp. 355–360, `https://doi.ieeecomputersociety.org/10.1109/BRACIS.2016.071`.
[15] Koplenig A., Meyer P., Wolfer S. and Müller-Spitzer C., *PLoS ONE*, **12** (2017) 1.
[16] Nettle D., *Philos. Trans. R. Soc. B: Biol. Sci.*, **367** (2012) 1829.
[17] Liu H. and Cong J., *Chin. Sci. Bull.*, **58** (2013) 1139.
[18] Bentz C., Alikaniotis D., Cysouw M. and Ferrer-i-Cancho R., *Entropy*, **19** (2017) 275.
[19] Shannon C. E., *Bell Syst. Tech. J.*, **27** (1948) 379.
[20] 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* (Neural information processing systems foundation) 2002.
[21] Hammarström H., Forkel R., Haspelmath M. and Bank S., *Glottolog 4.3* (Jena) 2020, `https://glottolog.org/accessed2021-04-08`.
[22] 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.

[23] Newman M. E. J. and Girvan M., *Phys. Rev. E*, **69** (2004) 026113.

[24] Blondel V. D., Guillaume J.-L., Lambiotte R. and Lefebvre E., *J. Stat. Mech.: Theory Exp.*, **2008** (2008) P10008.

[25] Traag V. A., Waltman L. and van Eck N. J., *Sci. Rep.*, **9** (2019) 5233.

[26] Newman M. E. J., *Phys. Rev. E*, **74** (2006) 036104.

[27] Traag V. A., Aldecoa R. and Delvenne J.-C., *Phys. Rev. E*, **92** (2015) 022816.

[28] Aldecoa R. and Marín I., *PLoS One*, **6** (2011) e24195.

[29] 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*, *ETMTNLP '02*, Vol. **1** (Association for Computational Linguistics, USA) 2002, pp. 63–70, https://doi.org/10.3115/1118108.1118117.

[30] 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.

[31] Rossetti G., Milli L. and Cazabet R., *Appl. Netw. Sci.*, **4** (2019) 52.

[32] Ferrer-i-Cancho R., Gómez-Rodríguez C., Esteban J. L. and Alemany-Puig L., https://arxiv.org/abs/2007.15342 (2020).

[33] Seabold S. and Perktold J., *Statsmodels: Econometric and statistical modeling with python*, presented at the *9th Python in Science Conference 2010*.

[34] Zipf G. K., *The Psychobiology of Language: An Introduction to Dynamic Philology* (Routledge, London) 1936.

[35] Zipf G. K., *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology* (Addison-Wesley Press) 1949.

[36] Altmann E. G. and Gerlach M., *Statistical Laws in Linguistics* (Springer International Publishing, Cham) 2016, pp. 7–26, https://doi.org/10.1007/978-3-319-24403-7_2.

[37] Petersen A. M., Tenenbaum J. N., Havlin S., Stanley H. E. and Perc M., *Sci. Rep.*, **2** (2012) 943.

[38] Tixier A., Malliaros F. and Vazirgiannis M., *A graph degeneracy-based approach to keyword extraction*, in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Austin, Tx., USA) 2016, pp. 1860–1870, https://www.aclweb.org/anthology/D16-1191.

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

[40] 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. Learn. Res.*, **12** (2011) 2825.

[41] Hastie T., Tibshirani R. and Friedman J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, *Springer Series in Statistics* (Springer) 2009, https://books.google.cl/books?id=eBSgoAEACAAJ.

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

[43] Bickel B. and Nichols J., *Fusion of selected inflectional formatives*, in *The World Atlas of Language Structures Online*, edited by Dryer M. S. and Haspelmath M. (Max Planck Institute for Evolutionary Anthropology, Leipzig) 2013, https://wals.info/chapter/20.

[44] Bickel B. and Nichols J., *Inflectional synthesis of the verb*, in *The World Atlas of Language Structures Online*, edited by Dryer M. S. and Haspelmath M. (Max Planck Institute for Evolutionary Anthropology, Leipzig) 2013, https://wals.info/chapter/22.