Sensing Language Relationships from Social Media

Diogo F. Pacheco[†], Priya Saha[†], Fernando B. de Lima-Neto[‡] and Ronaldo Menezes[†]
[†]BioComplex Laboratory, School of Computing, Florida Institute of Technology, Melbourne, FL, USA
[‡]Escola Politécnica, University of Pernambuco, Recife, Brazil
dpacheco@biocomplexlab.org, psaha@biocomplexlab.org, fbln@ecomp.poli.br, rmenezes@cs.fit.edu

***Abstract*—Online social networks (e.g., Twitter) offer an open platform for people to interact and connect without restrictions of language usage or geographic borders. Because of their pervasiveness, online social networks provide data and become real-time sensors of society. This work looks at Twitter to reveal the hidden relationship of languages that stems from users' language preference for writing their tweets. We show that the language relationships are dependent of place by comparing 12 large-scale datasets with different locality levels. For instance, the secondary language of French speakers in Canada is different from French speakers in France. We used network science and clustering techniques to find that languages groups are more driven by spatial than syntactic proximity. The characterization of language relationships is key to the understanding of information spread in social media and the detection of cultural shifts.**

I. Introduction and Motivation

The effect of globalization over the past few years has been observed in various domains of our lives, including trading, immigration, education, and culture [1]–[4]. Due to the connectedness of society, information (fads, trends, etc.) tends to be transmitted through people's social networks. This effect on society has been extensively discussed in the literature [5]–[7]. Yet, the impact of language to information spread has received very little attention. Today, a handful of languages have become globally popular; the popularization of TV in the late 50s and early 60s, as well as the current explosion in the use of social media, have all contributed to the popularity of certain languages. Although historic relationships among the languages are useful, understanding the significance of the language relationships from population preferences can lead to the identification of possible culture shifts.

Languages may be organized hierarchically according to historical relationships leading to language family trees; examples of a few popular language families are: Afro-Asiatic, Dravidian, Indo-European, Tai-Kadai, and Uralic. In a language tree, the closer the languages are from each other, the more similar they tend to be syntactically. Languages are not static, but evolve with society; for instance, Greek, Arabic, and Latin used to be popular but today English is considered the *de facto* global language. If one wants to study language relationships in today's world, the analyses have to consider social media, given its wide use. Social media has become the standard form of communication for the younger generation, as it provides an easy way for them to express their opinions [8].

In spite of several works related to the language of users in Twitter [9]–[12], the research community did not pay enough attention to the importance of *language relationships* generated from the user preference. Our work explores the characterization of languages as an emergent effect of individual online behavior.

First, we propose a user characterization based on the frequency of the languages used in his/her tweets. Then, we aggregate users based on their most used language. We use unsupervised machine learning algorithms and network science to understand the extent to which languages group together due to their origins (family trees) or other factors, such as geographical proximity. We use 12 different Twitter datasets (Table I) to understand and to capture language singularities.

This work reveals the structure of languages on Twitter and provides a window into how these languages are related to one another in this social media platform. Moreover, we demonstrated how a language relationships can differ from place to place and how to obtain a global characterization. The characterization we provide can be used to improve target campaigns, marketing strategies, or social network interventions. Potentially, it can be used as a tool to identify unexpected migrations. Finally, we provide some insightful visualizations on the structure of language relationships.

II. Methods and Technical Solutions

In this section, we describe how to characterize Twitter users based on the languages in which they tweet and how to aggregate the users in order to characterize the languages themselves; also, we describe the datasets used in this work based on their locality level. The aforementioned characterization is an adaptation from the work of Pacheco et al. [13].

A. Characterizing Users

We begin by characterizing users as the frequency of languages they use on their posts (tweets). Although a tweet can have words from different languages, in our datasets we only consider single-language tweets automatically detected by Twitter [14].

Let $\mathcal{T} = \{\tau_1, \ldots, \tau_m\}$ be the set of posts sent by m users in n languages, where τ_i are the posts sent by user i. We use these data to calculate a contingency table (i.e., a two-way table) of frequencies of languages per user. To ensure all users are treated equally regardless the number of posts they send, we normalize the rows of the matrix so that the elements of each row sum up to 1. The normalized contingency matrix of m users and n languages is $\hat{U} = [\hat{u}_{ij}]_{m\times n}$, where an element

978-1-5386-0435-9/17/$31.00 ©2017 IEEE

TABLE I

Descriptive statistics of 12 datasets identified by their locality level, global (◇), country(†), and city(⋆); the number of tweets and users; the total number of languages $|L|$ and languages used by at least 10 users $|L|^+$; the percentage of monolingual users; the average number of languages used by multilingual users; and the collection period.

Dataset	Tweets	Users	$\|L\|$	$\|L\|^+$	% Mono.	$^L/_{\text{User}}$	From	To
2016 Olympic Games◇	18,048,522	6,506,634	61	55	93%	2.17	08/01/16	08/24/16
G20◇	10,610,653	2,694,784	60	50	93%	2.28	08/24/14	09/29/14
2015 Women's World Cup & America Cup◇	10,026,573	2,704,898	62	48	90%	2.38	06/16/15	07/13/15
2014 FIFA World Cup◇	50,476,375	9,235,153	64	48	73%	3.10	06/12/14	07/13/14
2016 UEFA Euro◇	36,456,419	5,413,895	60	48	79%	2.78	06/10/16	07/19/16
The United Kingdom†	30,373,072	3,069,664	65	51	84%	2.71	02/07/15	05/07/15
South America†	334,337,906	2,743,842	66	44	56%	5.71	04/23/15	12/08/15
New York City⋆	1,925,831	130,368	55	38	80%	2.92	08/29/14	09/29/14
Paris⋆	434,969	30,324	46	34	64%	3.18	03/09/15	04/03/15
San Francisco⋆	717,555	62,989	49	33	82%	2.77	03/05/15	05/05/15
Tokyo⋆	2,153,586	147,140	57	34	92%	2.55	03/05/15	05/05/15
Hong Kong⋆	96,302	10,682	43	23	61%	2.93	03/05/15	05/06/15

$\hat{u}_{ij}$ is the number of tweets from user i in language j divided by the total number of his/her tweets and it is given by

$$\hat{u}_{ij} = \frac{1}{|\tau_i|}\sum_{t\in\tau_i}\delta(t_\ell, j), \tag{1}$$

where $\delta(t_\ell, j) = 1$ if the tweet language $t_\ell = j$.

A user is characterized by a distribution of language probabilities, i.e. rows of the normalized contingency matrix $\hat{U}$. Therefore, the number of users using a language L_j is the sum of the partial contribution of all m users to language j,

$$L_j = \sum_{i=1}^{m}\hat{u}_{ij}. \tag{2}$$

B. Characterizing Languages

This work proposes to define languages based on users, which in turn, are also defined by languages. So, to characterize a language by its relations to other languages, one might assume each user has a *preferred* language. A user's preferred language is the one he/she tweets the most. Then, we encode this user-language preference in the membership-indicator matrix $P = [p_{ij}]_{m\times n}$ by conditioning i^{th} user $\in j^{th}$ language-group, such as:

$$p_{ij} = \begin{cases} 1 & \text{if } j = \arg_j \max \hat{U}(i,j) \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Next, we define a $n \times n$ language characterization matrix $K = [k_{jw}]_{n\times n}$, where the j^{th} row is the mean probabilities of all users whose preferred language is j. Consequently, a language is characterized relative to all languages as a distribution of probabilities, i.e., it encompasses the average behavior of its preferred "speakers" (users). Languages (rows of matrix K) are given by

$$\mathbf{k}_j = \frac{\sum\limits_{i=1}^{m} p_{ij}\hat{\mathbf{u}}_i}{\sum\limits_{i=1}^{m} p_{ij}}, \tag{4}$$

where $\hat{\mathbf{u}}_i$ is a row in user matrix $\hat{U}$, and p_{ij} is the flag in the membership-indicator matrix P indicating whether it belongs to the language-group.

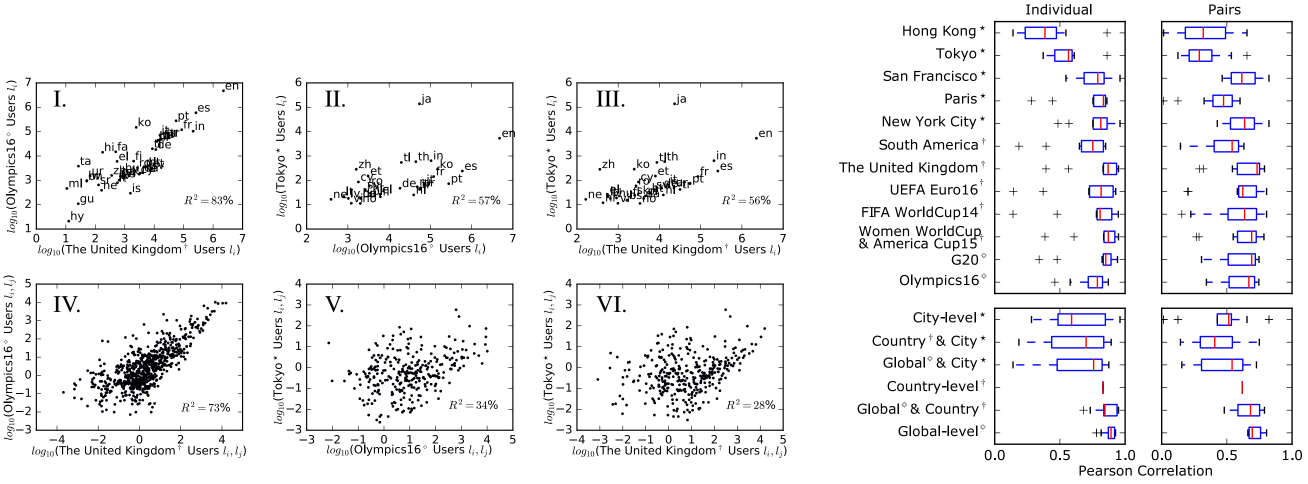
C. Data

The datasets used here (see Table I) vary in the collection process (common terms or bounding boxes), yielding different levels of locality. The datasets that were collected by tracking terms are event-based and, therefore, global. The other datasets, collected using bounding boxes, are constrained to geographical boundaries, and consequently, they are limited to cities and countries within the box.

Table I shows some statistics of the datasets. Data was gathered within a 3-year period. The datasets vary in number of tweets (0.1–334 million) and users (0.01–9 million). The number of languages used by multilingual users is quite stable among all datasets. However, the number of languages used by at least 10 users $|L^+|$ on global-level datasets is larger than on city-level, suggesting distinct features among them.

Different places are formed by an amalgam of different cultures. Therefore, it is expected that the languages characterized by these datasets generate dissimilar networks and clusters as they embed the peculiarities of different places. To verify the similarities among these independent datasets, Figures 1(a) and 1(b) compare the correlation of the number of users using each language individually (Equation 2) as well as those using pairs of languages (the number of users tweeting in Japanese and Portuguese, for example); this validation process was proposed by Ronen et al. [11].

Figure 1 shows the results of the datasets validation. The comparison example using 3 datasets in Figure 1(a) and the overall evaluation in Figure 1(b) suggest datasets tend to be more similar as their locality level decreases. In other words, the relationship between pairs of languages in global datasets tend to be more similar than when compared against the relationship in a city-level dataset. Hence, global datasets are more adequate to describe languages while city ones should



(a)

(b)

Fig. 1. Datasets validation. (a) Comparing the correlations between pairs of datasets at different levels of locality: global (Olympics16), countries (The United Kingdom), and city (Tokyo). The correlations are higher for the pairs with the global dataset than for the pairs with the city one. The top row shows the correlation between the number of users tweeting in a language, i.e. each point is a specific language; while the bottom row shows the correlation of users with tweets in pairs of languages, e.g. the number of users using English and Japanese. Comparison between datasets: Olympics16 and The United Kingdom (I and IV); Olympics16 and Tokyo (II and V); and The United Kingdom and Tokyo (III and VI). (b) Measuring the locality effect of datasets on language relationships. Each boxplot contains the Pearson correlations for 11 pairs of datasets. From *Hong Kong* (top) to *Olympics16* (bottom), datasets vary from more local to more global level as depicted by the following special marks: city-level (⋆), country-level (†), and global-level (◇). On average, global datasets presents higher correlation than city datasets. On the left, correlations between users tweeting in one language, while on the right, the correlations of users tweeting in pairs of language. The top blocks show individual dataset comparisons while bottom blocks show inter and intra-levels comparisons.

TABLE II
TOP EIGHT LANGUAGES RANKED IN THE LN BASED ON *in/out* DEGREE AND WEIGHTED DEGREE, AND EIGENVECTOR CENTRALITIES.

In-Deg	W. In-Deg	Out-Deg	W. Out-Deg	Eigenvector
English	English	English	Indonesian	English
Indonesian	Russian	Spanish	Armenian	Indonesian
Finnish	Indonesian	Portuguese	Marathi	Spanish
Spanish	Persian	Indonesian	Oriya	Finnish
Estonian	Spanish	Italian	Gujarati	Portuguese
Portuguese	Hindi	French	Hindi	Italian
Tagalog	Tagalog	German	Tamil	French
Italian	Portuguese	Russian	Serbian	Tagalog

be used to discover language singularities within regions. Due to space limitations, this paper focuses on the analysis on the most global dataset – Olympics16. Unless explicitly mentioned otherwise, all figures are based on this dataset.

III. Empirical Evaluation

We propose to use K as an adjacency matrix to generate a weighted-directed network of languages. In addition, we represent each $\mathbf{k}_j$ as feature vectors, and use a hierarchical clustering algorithm to identify similarities among languages.

A. Language Network

The language characterization matrix K (from Equation 4) can be used as an adjacency matrix to build a language network (LN). When K is generated from a global dataset, such as the Olympics16, there is a global LN. The LNs are weighted-directed networks, where nodes are languages, and edge-weights represent the proportion of users interacting in both the source and target languages. For instance, a link of 0.02 from Dutch to English means that Dutch users write in English in 2% of their social interactions. The self-edges are ignored in the LN since our interest lie in understanding the relationships among languages.

Table II presents the top 8 languages based on a few centralities. The *in-degree* represents the diversity of the neighborhood of a language (i.e. how many languages have at least one user who also uses the language in question), while the *weighted in-degree* of a language captures the volume of speakers for each of these connections captured by the in-degree. For instance, Italian is among the top 8 languages with high *in-degree*, while Russian is among the top 8 languages with high *weighted in-degree*; therefore, people from different cultures use more Italian than Russian, but the fewer cultures who use Russian, do it much more frequently. The *out-degree* of a language demonstrates the *multilingualism* of the users of the language or the tendency of users to connect to others with distinct preferred languages. In the LN, language relationships are asymmetric; while in-degree can be seen as a measurement of language popularity, the out-degree can be seen as a plurality indicator of users of a particular language. Finally, the *eigenvector* is an influential centrality (diversity and volume) since it considers the structure of the network.

Figure 3-A shows the language network. The sizes of nodes represent the weighted in-degree, and the colors are the communities according to Blondel's algorithm. We used language community colors to set countries' color in the

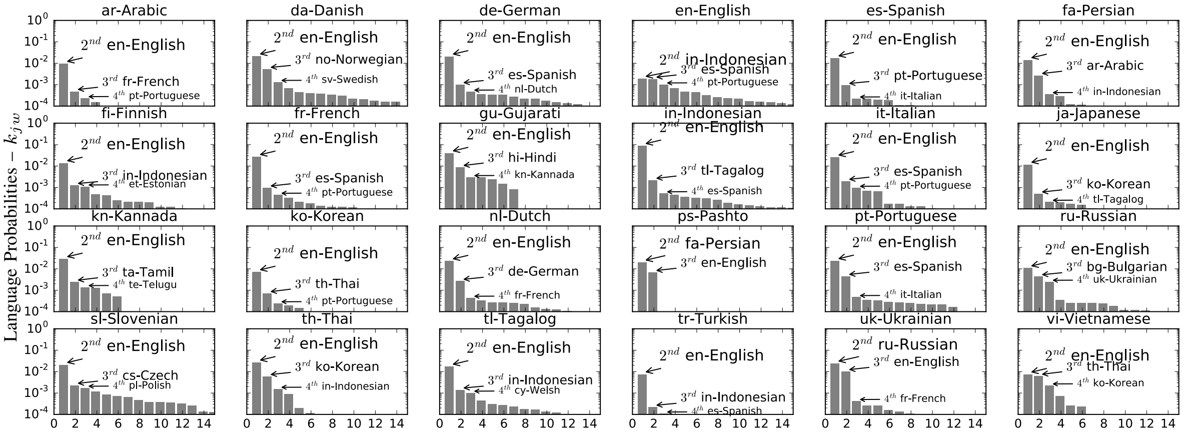


Fig. 2. A sample of language characterizations from Olympics16 dataset. Probabilities (k_{jw}) are in log scale and the three most used languages other than itself are highlighted; e.g., among nl-Dutch speakers, around 2%, 0.3%, and 0.04% also tweet in English, German, and French.

world map on Figure 3-F. Each country assumes the color of its major language according to the CIA's World Factbook[1]. For instance, the language distribution in Canada is 59% English, 22% French, etc., so English is its major language. Consequently, Canada is yellow in the map since English is in the yellow community in the LN. India is colored by states, since it has multiple major languages (defined at state level[2]), and they are located in different communities. The map shows neighboring areas (contiguous) with predominantly same colors, suggesting that the LN embeds some notion of geographical proximity.

B. Language Characterizations

The language network gives the overview of language interactions. In order to get a more specific understanding about a language, we can use a visual representation of it. Plots of the rows of matrix K (language probability distributions) are very informative, specially when presented as ranked histograms.

A sample of language characterizations is shown in Figure 2. For instance, among Dutch users[3], around 2%, 0.3%, and 0.04% also tweet in English, German, and French, respectively. English is the second preferred language for the majority of users of all languages, except for Bulgarian, Pashto, and Ukrainian users. Clearly, there are distinct levels of *multilingualism* among languages; for example, Pashto users tend to, collectively, communicate in at most 3 languages, while English users collectively use more than 14. The "shape" of the language distributions can be used to explore these differences.

One can notice some similarities, even by highlighting only the 3-most used languages per plot. For instance, in Spanish and Portuguese, the 2^{nd} and 4^{th} languages are English and Italian, respectively, while they are both the 3^{rd} option for each other. Moreover, the distribution of the languages are not always symmetric, such as Spanish–Portuguese or Thai–Korean. German is more important to Dutch users than Dutch is for German users. Similarly, Ukrainian users prioritize more Russian than the other way round.

C. Language Clusters

The language distributions reveal more details about individual languages than the language network; they can more easily show similarities among languages. We investigated the extent of similarity among the languages by using the Agglomerative Clustering algorithm. The hierarchical clusters can provide additional information between the specificity of a language distribution and the generality of a language network. The elements to be clustered are languages (rows of matrix K), where each component represents the probability that a user of a language communicates using another language. We used the Bhattacharyya distance as the affinity (distance) metric, since it is more suitable for comparison of discrete probability distributions than other metrics, such as Euclidean distance [15]. We also tested the Cophenetic Correlation Coefficient and determined that "average" (the UPGMA algorithm) was the best linkage method [16].

Figures 3-B to 3-D depict the hierarchical cluster for three datasets: Tokyo, the United Kingdom, and the Olympics16. Despite the differences between the datasets, some clusters remain consistent, such as Portuguese–Spanish; Chinese–Japanese; Norwegian–Danish–Swedish; and Dutch–German. However, to have a more general representation of language similarities, rather than local idiosyncrasies, we focus on the clustering of the Olympics16 dataset (Figure 3-D).

As for the LN, Figure 3-G shows a world map where countries are colored based on clusters of languages as defined in Figure 3-D. Upon close inspection, the map shows neighboring countries belonging to the same cluster, equivalent to the contiguous coloring pattern obtained from communities (Figure 3-F); the spatial correlation is observed in both analysis.

The visual effect of the spatial correlation is barely changed, regardless of the fact that clusters tend to be smaller than

[1] https://goo.gl/H5wplB

[2] https://goo.gl/7mQzu6 and https://goo.gl/JW6jDz

[3] In this context, Dutch, Italian, French, etc. refer to the preferred language rather than nationality.

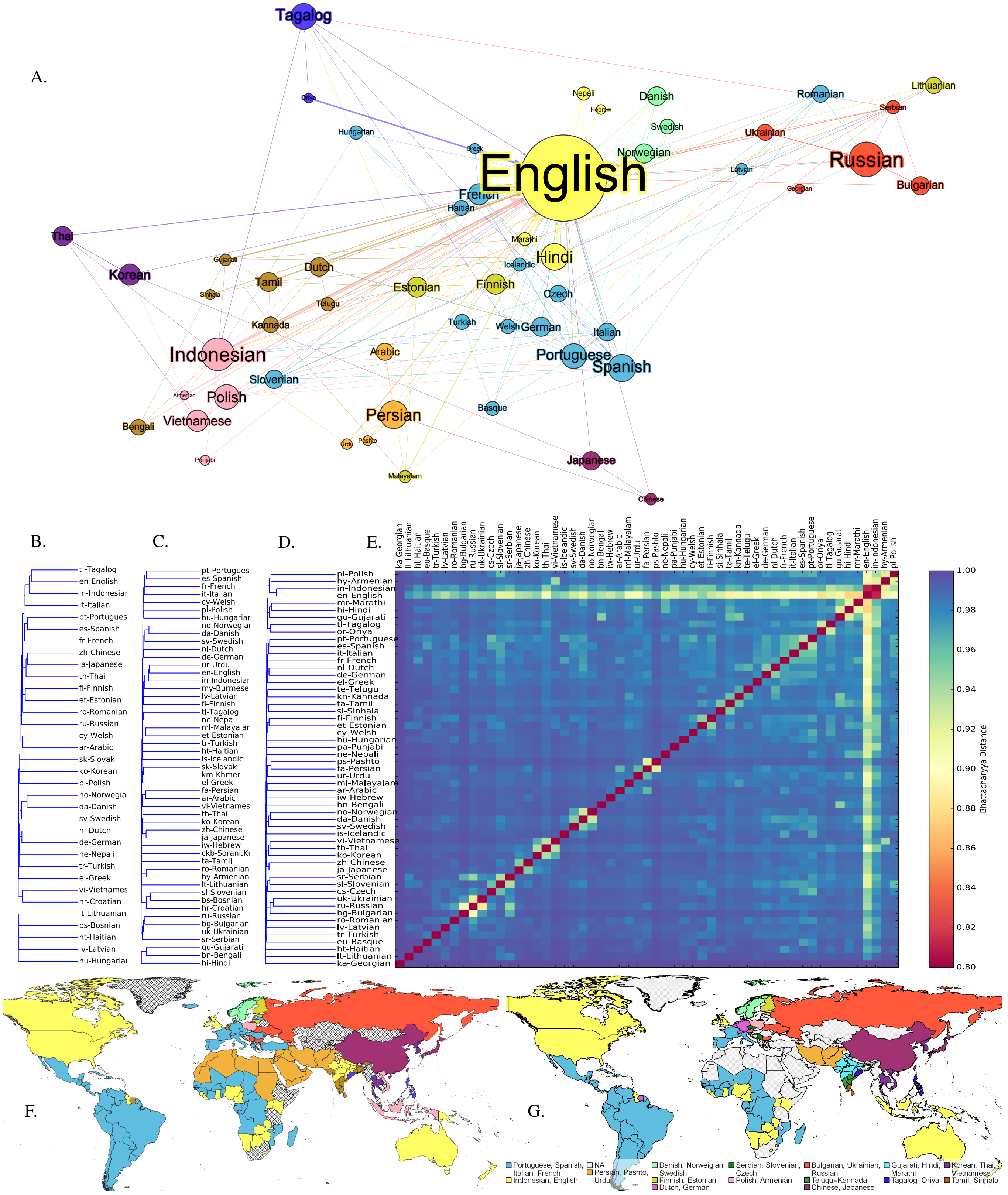


Fig. 3. Revealing language relationships. Except for (B–C), plots are based on the Olympics16 dataset. (A) The language network – nodes are languages, colored based on their communities, and sized by the weighted in-degree; directed links colored as the source node representing the percentage of users in the source language that also use the target language. (B–D) The language hierarchical clusters for datasets at different locality-levels: (B) Tokyo, (C) The UK, and (D) Olympics. (E) The similarity matrix used to create the hierarchical cluster (D). (F–G) The spatial dependence captured when grouping languages; countries are colored according to the group of their major language: (F) a community in the LN or (G) a cluster. Groups tend to fill contiguous neighboring.

communities. For instance, although Telugu–Kannada and Tamil–Sinhala are two different clusters, all four languages are adjacent in the map and also belong to the same community on the LN. Yet differences are present between the two approaches, for instance German–Dutch and Czech–Slovenian–Serbian share clusters but in the community approach they are not together, even though the countries where the language is mainly spoken share some geographical proximity.

The results showed that languages are not grouped in communities or clusters based solely on syntactic similarity (family trees). Groups are composed by languages widely spread in the language family tree structure, such as: the different-family group formed by Chinese (Sino-Tibetan), Japanese (Japonic), Korean (Koreanic), Thai (Tai-Kadai), and Vietnamese (Austroasiatic); or the same-family-different-branch group formed by Indo-European languages such as Persian and Pashto (Iranian branch) and Urdu (Indo-Aryan branch).

The syntactic diversity of languages found in groups by independent methods (communities and clustering) and the consistent spatial correlation patterns shown in the maps, suggest that other factors beyond language family proximity drive people in their choices for interaction and social dependencies.

IV. Conclusion

In this work, we classified languages based on their relationship with other languages. Each language is represented as the emergent behavior from its *preferred* users. Users, on the other hand, are characterized based on the frequency of languages used in their tweets. We used 12 different datasets to understand whether the locality of a language had impact on its characterization.

Indeed, our results suggest that the relationship between languages is not limited to their origins (language family-tree), but is strongly dependent on spatial factors (such as the sharing of borders). The support for this argument is two-fold. First, local datasets do not correlate to each other; if family branches were dominant they should correlate. Second, the contiguous groups of adjacent countries are quite similar for both techniques used in the classification (clustering or community detection).

The characterization presented here has limitations since we cannot overcome possible bias embedded in Twitter data [17]. However, it is worthwhile noting that our characterization is centered on individuals, regardless of them being multilingual or their tweets being geo-tagged. The language characterization itself is a consequence of how we choose to aggregate users. Consequently, the results presented here tend to be less biased than those whose characterization considers multilingual users only [10], [11]. Yet, we plan to evaluate the precision of the Twitter language detection in our datasets by comparing the results against other tools, such as the chromium compact language detector.

We plan to do a characterization of places (based on their languages). This would require a simple redefinition of the membership function (Equation 4); we expect to have results on this approach in the very near future.

We believe our approach can be used to sense populations in real time, possibly detecting abrupt cultural shifts, such as in the presence of massive displacement of refugees. More importantly, the approach allows us to understand language barriers formed in social media. This can be useful when working with information spread applications (e.g. marketing campaigns, social network interventions). Our approach appears to indicate that information is likely to spread in accordance to spatial location of the countries more than based on language family branches.

References

[1] Y.-A. de Montjoye, L. Radaelli, V. K. Singh, and A. S. Pentland, "Unique in the shopping mall: On the reidentifiability of credit card metadata," *Science*, vol. 347, no. 6221, pp. 536–539, jan 2015.

[2] M. Peres, H. Xu, and G. Wu, "Community evolution in international migration top1 networks," *PLoS ONE*, vol. 11, no. 2, p. e0148615, feb 2016.

[3] J. D. Hansen and J. Reich, "Democratizing education? Examining access and usage patterns in massive open online courses," *Science*, vol. 350, no. 6265, pp. 1245–1248, dec 2015.

[4] A. Teymoori, J. Jetten, B. Bastian, A. Ariyanto, F. Autin, and et. al., "Revisiting the measurement of anomie," *PLoS ONE*, vol. 11, no. 7, p. e0158370, jul 2016.

[5] R. Lambiotte and M. Kosinski, "Tracking the digital footprints of personality," *Proceedings of the IEEE*, vol. 102, no. 12, pp. 1934–1939, 12 2014.

[6] S. A. Golder and M. W. Macy, "Digital Footprints: Opportunities and Challenges for Online Social Research," *Annu. Rev. Sociol*, vol. 40, no. May, pp. 129–52, 7 2014.

[7] C. E. Tucker, "Social networks, personalized advertising, and privacy controls." *Journal of Marketing Research*, vol. 51, no. 5, pp. 546–562, 2014.

[8] S. Gearhart and W. Zhang, "was it something i said?no, it was something you posted! a study of the spiral of silence theory in social media contexts," *Cyberpsychology, Behavior, and Social Networking*, vol. 18, no. 4, pp. 208–213, 2015.

[9] B. Poblete, R. Garcia, M. Mendoza, and A. Jaimes, "Do All Birds Tweet the Same ? Characterizing Twitter Around the World," *Society*, pp. 1025–1030, 2011.

[10] D. Mocanu, A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, and A. Vespignani, "The Twitter of Babel: Mapping World Languages through Microblogging Platforms," *PLoS ONE*, vol. 8, no. 4, p. e61981, 4 2013.

[11] S. Ronen, B. Gonçalves, H. K. Z, A. Vespignani, S. Pinker, and C. A. Hidalgo, "Links that speak: The global language network and its association with global fame," *Proceedings of the National Academy of Sciences*, vol. 111, no. 52, p. 201410931, 12 2014.

[12] P. Saha and R. Menezes, "Exploring the World Languages in Twitter," in *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI' 16)*, 2016, pp. 153–160.

[13] D. F. Pacheco, D. Pinheiro, F. B. Lima-Neto, E. Ribeiro, and R. Menezes, "Characterization of Football Supporters from Twitter Conversations," in *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI' 16)*, 2016, pp. 169–176.

[14] B. Pavliy and J. Lewis, "The Performance of Twitter's Language Detection Algorithm and Google's Compact Language Detector on Language Detection in Ukrainian and Russian Tweets," *Bulletin of Toyama University of International Studies*, vol. 8, pp. 99–106, 2016.

[15] T. Kailath, "The Divergence and Bhattacharyya Distance Measures in Signal Selection," *IEEE Transactions on Communication Technology*, vol. 15, no. 1, pp. 52–60, 2 1967.

[16] R. R. Sokal and F. J. Rohlf, "The Comparison of Dendrograms by Objective Methods," *Taxon*, vol. 11, no. 2, pp. 33–40, 2 1962.

[17] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist, "Understanding the demographics of twitter users." *ICWSM*, vol. 11, p. 5th, 2011.