Language as a Measure of Welfare

Priya Saha, Diogo Pacheco and Ronaldo Menezes
BioComplex Laboratory, School of Computing
Florida Institute of Technology, Melbourne, USA
{psaha,dpacheco}@biocomplexlab.org, rmenezes@cs.fit.edu

Abstract

Globalization is a process driven by international trade that leads to interactions and integration of people and government of different nations. Such process has been impacting us in many ways including our living standards or quality of life (QoL). At the same time, the integration between peoples of the world leads to a stronger diffusion of languages, ideas, and values; more recently the integration has received further boost by the emergence of online social networks. Online social networks give us a platform to connect without any restriction to geographic regions, language usage, costumes, etc. In fact, they are the quintessential example of globalization. Yet the links between language usage in society and QoL has received little attention. Since both language usage and the QoL are influenced by globalization it is just natural that one tries to study both subjects combined. This paper investigates if one can be used as a proxy for the other. Using approaches based on network science, our analysis of a large-scale Twitter dataset reveals that the patterns of user connectivities on online social networks (such as Twitter) as a function of languages usage is correlated to the QoL.

Introduction and Motivation

Evidence of globalization has been observed in many domains such as culture, economy, and international policies. The expansion of international trade and foreign investment is often seen as a sign of economic growth; higher economic growth rates and greater affluence are conducive to wellbeing (as in QoL). However, what is the wellbeing of a society? Many critics argue that GDP cannot capture the nation's of wellbeing (Milenkovic et al. 2014). They claim that GDP is intended to measure the *productivity* of a nation and hence is an insufficient measure to quantify the QoL; wellbeing of a nation is a multi-dimensional concept and not the economic growth alone. There could be other social conditions that play a role in improving QoL. On the other hand, *linguistic imperialism* or the forcing of the dominant language in today's world is also an outcome of (cultural) globalization. This paper looks at how these two concepts (language and QoL) are intertwined; we are interested in understanding the extent to which language can be seen as a proxy of wellbeing.

The United Nations (UN) introduced a measure called Human Development Index (HDI) as a shift away from economic growth as the only measure of prosperity. The HDI is a composite metric that considers life expectancy, education and per-capita income in its equation, and is consequently a better indicator of QoL of people than GDP. The HDI allow countries to be ranked in four tiers: low, medium, high and very high. Although GDP is sometimes used as an indicator of prosperity, it does not necessarily correlate to HDI. An example is Cuba which according to the 2007 data, has low GDP of PPP US$6,876 and very high HDI of 0.863 (the max is 1) (Sandrin 2011).

Authors have tried to look at relations between QoL and other factors. Ranis found that human development has important effects on economic growth (**?**). An increase in the capabilities available to individuals allow them to pursue occupations in which they are more productive. In this sense, human development is correlated to the human capital and human capital, in turn, correlates with the economic growth. Several studies used HDI in an attempt to understand specific human characteristics. One of the more interesting examples look at the relation between HDI and obesity levels (McLaren 2007). Mocanu et. al characterized the worldwide linguistic diversity in Twitter using geotagged data at different scales from country to neighborhood scales (Mocanu et al. 2013). They show that the usage of Twitter is not uniform and has a correlation with economical factors. In another study, Kulshrestha et al. demonstrated the influence of geography in the cultural and linguistic backgrounds in Twitter (Kulshrestha et al. 2012). Economic imbalances in society is correlated to the imbalance in the total number of tweets. For example, US accounts for 25% world GDP and 72% of all tweets produced in Twitter. Ronen et al. studied the interactions between languages using one billion tweets, book translations and editors of Wikipedia (Ronen et al. 2014). English was found as a hub in the network. This study revealed that there exists a strong correlation between the number of fa-

mous people native to a language and the position (eigenvector centrality) of the language in the network. Saha and Menezes also demonstrated that the positions of languages on Twitter indicate several interesting insights including the visibility of information generated in a particular language (Saha and Menezes 2016a; 2016b).

Although the observation of the connection between language and economic growth is interesting, what is more intriguing is to see if languages used by people on Twitter can act as a proxy to their development scale. Twitter data has gained a lot of interest among the research scientists who are trying to understand the specific aspects of human behavior because it can act as a large real-time sensor of society (Demirbas et al. 2010)

Our main contribution in the paper is to show the relation between the position of the languages in networks generated from datasets extracted from Twitter and the human development of the language (calculated as the function of the countries using that language). We validated our findings using statistical methods. The rest of the paper is organized as follows: we discuss our procedure for collecting data and generating language networks, followed by a discussion on calculating the HDI of languages. Next, we describe our experiments and finish the paper by discussing our views and conclusions.

Data Collection and Network Generation

We used 2 Twitter datasets to analyze the correlation of languages and the HDI. We used the language of the tweet as detected by Twitter (the language is available as metadata of the tweet). The datasets were collected at a global level to avoid language bias. A global level dataset can capture much more language diversity in the network. Below, we describe the dataset we collected and used to generate language networks (see Table 1 for full details of the datasets.

Dataset 1: G20

Our first dataset is a collection of about 10 million tweets posted for a period of 35 days about the leaders of the Group of Twenty[1]. The G-20 involves 19 individual countries plus the European Union. We collected the tweets that consisted of one or more of the last names of the leaders of the G-20. By capturing the tweets about the G20 leaders, we were able to capture a great deal of language diversity in Twitter.

Dataset 2 : Olympic Games 2016

The Olympics Games 2016 dataset consists of about 18 million tweets collected by using the keyword *olympics* in several languages. The summer games was hosted in Brazil with more than 11,000 athletes from 207 countries. The keywords are demonstrated in Figure 1. We used the Google Translator to get translations of the "Olympics" term.

[1]https://www.g20.org/

Table 1: Descriptive statistics about the datasets. Monolingual refers to users who use one language and multilingual refers to users who use more than one languages to tweet.

Statistic	G20	Olympics 2016
Start Data Collection	August 24, 2014	August 01, 2016
Finish Data Collection	September 29, 2014	August 24, 2016
Number of Days	35	24
Tweets with Identified Language	10,610,653	18,048,522
Number of Users	2,694,784	6,506,634
Number of Languages	55	50
Monolingual Users	93%	93%
Avg. Languages / Multilingual User	2.28	2.17

For each user, we collected the languages she used to tweet. Then we aggregated all the users who demonstrated a particular language as their most-frequent language. In the language network, we connect the frequent language and the other common languages with a link having an edge weight of the average value for all users who prefer a particular language but also uses another language. For example, say user A tweets 80% in English and 20% in Spanish. So the language vector of A can be represented as $A_{\text{Eng,Spa}} = [0.2]$. Similarly, we generate the language vectors of other users who prefer English as their frequent language and also tweet in Spanish. Say, $B_{\text{Eng,Spa}} = [0.3]$, $C_{\text{Eng,Spa}} = [0.4]$. In the language network, we connect English and Spanish with a link of edge weight 0.3 (the average). We repeated the same process for every language combinations and generated language networks from both the datasets. The language networks are directed.

HDI of Languages

There can be many multilingual people in a country who communicate in more than one language and the distribution of the language users are not evenly distributed. We collected the Human Development Index of every country as reported by United Nations, the language distribution in every country data in (Ronen et al. 2014), and the percentage of speakers of every country for every language as reported in the World Factbook by the Central Intelligence Agency(Agency 2016). Next, we computed the HDI of every language by the weighted average below

$$\text{HDI}_\ell = \frac{\sum_c (H_c N_{\ell c})}{N_\ell}, \tag{1}$$

where HDI_ℓ is the Human Development Index of language ℓ, H_c is the Human Development Index of a country c, $N_{\ell c}$ is the number of speakers of ℓ in country c, and N_ℓ and the total number of speaker of language ℓ in the world. The HDI of the languages are approximate because the values depend on many different factors. The HDI of a language is the average contribution of a

Figure 1: The term "Olympics" was translated to several languages and used to collect tweets related to the Summer Olympic games in Rio de Janeiro, Brazil.

single speaker of the language towards the world HDI, therefore summing the contributions of all the speakers of the language to the HDI of every country and then dividing by the total number of speakers. Our calculations are based on the data available from the Central Intelligence Agency and the United Nations.

Experimental Results

After generating the language networks, we performed multiple analysis to understand the inter-relations of the languages in Twitter. We start our analysis with some fundamental concepts of network science. Though the concepts are simple, we explore them to uncover the important characteristics described later in this paper.

Total Users by Language

The language distribution of the users can show us the tweet activities in general. Since Twitter has no restriction on how users can tweet (except that the tweets have to be less than 140 characters), the effects are directly reflected in the activities. In many cases, a user may tweet in more than one language. Figure 2 and Figure 3 show the contribution of the users to the languages in our datasets.

According Figures 2 and 3, English has the largest demographic followed by Spanish. Although the ranks of the languages are not the same, the top languages are similar in both the datasets. It is important to note that our rankings of the users by language do not reflect the estimates of the world speakers by languages. Ethnologue reports that Chinese is the most spoken language followed by Spanish and English (Grimes et al. 1988). There are several factors that may influence the different results: the penetration of Twitter in the population depends of age and census composition of the users. However, the disparities do not hinder in extracting interesting insights about the speakers in countries where Twitter is popularly used. Since our data considers the distribution of language users in different parts of the world, our findings are relevant and reflect the way users tend to interact on Twitter. In fact, it is a reminder to us that we are trying to derive language characteristics of the users using only Twitter data (Mocanu et al. 2013).

Languages by HDI

In order to analyze the relatedness of the languages and the development of the countries where they are spoken, we used Equation 1 to calculate the HDI_ℓ of all the languages that are present in our datasets.

Table 2: The 5 languages with highest and lowest HDI_ℓ.

High HDI_ℓ		Low HDI_ℓ	
Language	HDI_ℓ	Language	HDI_ℓ
Norwegian	0.94	Haitian	0.48
German	0.91	Punjabi	0.55
Dutch	0.91	Nepali	0.55
Swedish	0.91	Khmer	0.56
Danish	0.91	Urdu	0.59

Table 2 shows a few languages with very high and very low HDI_ℓ. Norwegian users display very high HDI_ℓ and Haitian users display low HDI_ℓ.

Table 3: The 5 countries with highest and lowest HDI_ℓ.

High HDI		Low HDI	
Country	HDI	Country	HDI
Norway	0.94	Nigeria	0.35
Australia	0.94	Central African Republic	0.35
Switzerland	0.93	Eritrea	0.39
Denmark	0.92	Chad	0.39
Netherlands	0.92	Burundi	0.40

Table 3 shows the HDI of the countries. Norwegian is dominantly spoken in Norway (which tops the HDI ranks for 12 years (Rebello 2015)). Hence, a Norwegian user is expected to have a better standard of living than a Haitian user. Haitian is dominantly spoken in Haiti.

Correlation with Network Properties

We used the centralities of the nodes (languages) in the language networks as the measure for their importance. The properties we considered for our analysis are in-degree, out-degree, betweenness, closeness, eigenvector centrality, weighted in-degree and weighted out-degree. Every network characteristic has its own interpretation and below we discuss the correlation analysis.

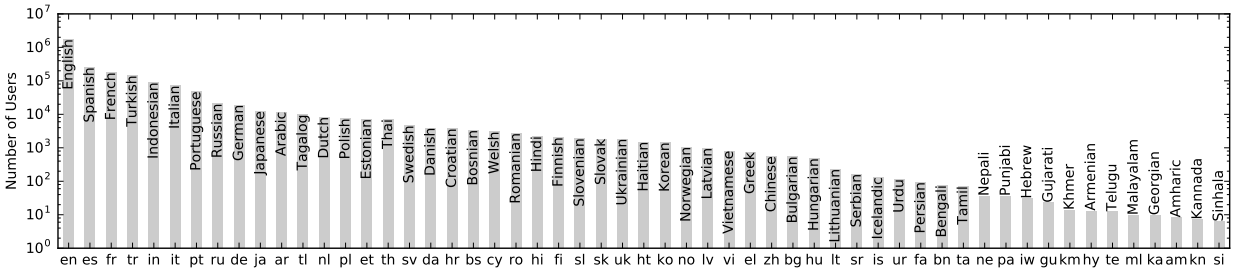


Figure 2: Distribution of users per language in the G20 dataset.

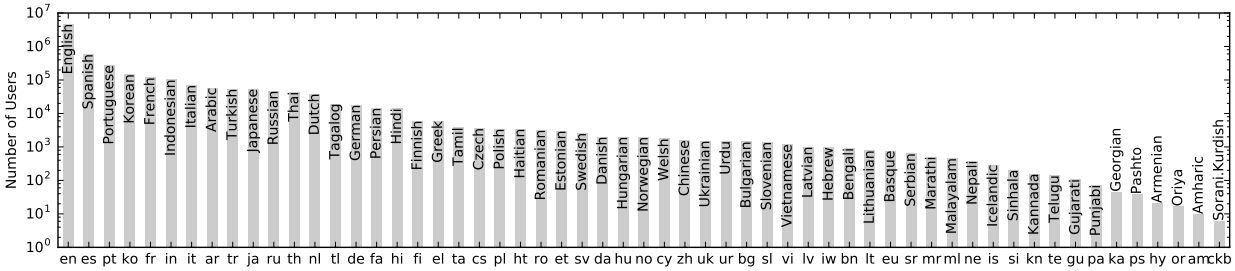


Figure 3: Distribution of users per language in the Summer Olympics dataset.

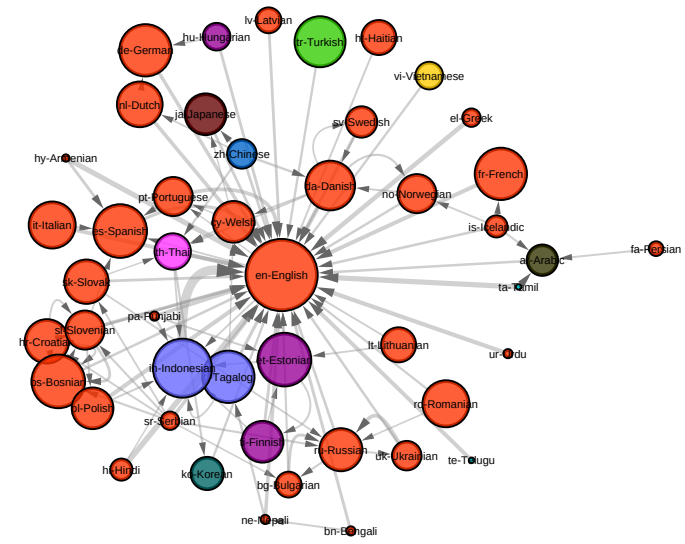


Figure 4: Language network of the G20 dataset: The color of a node represents the language family and the size of the node represents the in-degree. The represented language families are: Turkic, Indo-European, Japonic, Tai-Kadai, Austronesian, Uralic, Koreanic, Austroasiatic, Sino-Tibetan, Afroasiatic, Dravidian.

Table 4: Rank of the top languages in the G20 dataset according to in-degree and eigenvector centrality

In-degree	Eigenvector Centrality
English	English
Indonesian	Indonesian
Bosnian	Spanish
Spanish	French
Estonian	Turkish

First, we collected the characteristics of the languages in both networks. We listed the top 5 languages according to the in-degree and eigenvector centralities in the G20 as well as the Olympics datasets in Table 4 and Table 5 respectively. Figure 4 and Figure 5 show the language networks of the G20 and the Olympics datasets respectively. We show that a few languages in Twitter receive a great deal of attention from the users. Users who use the popular languages are in advantageous positions in the networks because they are likely to have more information available to them. We also observe that the Indo-European language family is very widely used in Twitter.

Next, we performed a correlation analysis between the network metrics and the HDI_ℓ. In both the G20 and the Olympics datasets, we found that the HDI_ℓ correlate significantly with the in-degree as well as the eigenvector centrality of the languages.

Table 5: Rank of the top languages in the Olympics dataset according to in-degree and eigenvector centrality

In-degree	Eigenvector Centrality
English	English
Indonesian	Indonesian
Finnish	Spanish
Spanish	Portuguese
Estonian	Italian

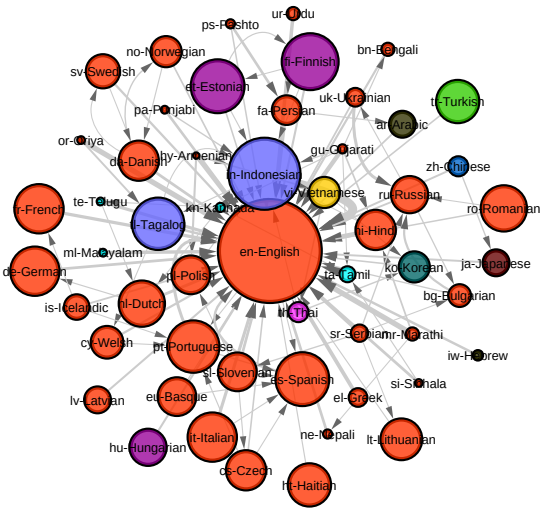


Figure 5: Language network of the Olympics dataset. The size and color represents a similar concept as Figure 4.

Eigenvector centrality considers the connectivity of a language as well as the connectivities of its neighbors in an iterative manner while in-degree of a language measures its incoming links. Languages that have high eigenvector centrality and in-degree have favored positions in the network. Our analysis shows that languages that hold such favored position in the networks, also have high HDI_ℓ. The users of such languages tend to have a better life.

Figure 6 demonstrates the correlation (described above) between the HDI_ℓ and their positions in the networks. In the in-degree vs HDI_ℓ analysis of both the datasets, we notice that a few popular languages such as Indonesian and French are below fitting line. Based on their prominence in the network, they were expected to have better HDI_ℓ. We observe that the standard deviation of the HDI of the Indonesian users in the different countries vary to a great extent. As a result, the variation influences the overall HDI_ℓ. French does not have high standard deviation. The low HDI_ℓ of French could be because of many different factors such as Twitter penetration in the countries where it is used or the population of the countries. On the other hand, languages such as Norwegian and Dutch are above the fitting line. Although Norwegian and Dutch are not as prominent as English in the network, they are spoken in countries having high HDI. There is a possibility that Norwegian and Dutch may gain much more prominence in the network. The fact that some languages are not very central in the network is not related to the HDI. We summarize the results with statistical significance in Table 6. We found that eigenvector centrality and the in-degree positively correlated with HDI with statistical significance in both the datasets. In Table 6 we indicated the first two rows as the ones that are statistically significant.

Table 6: The correlation of the network metrics with HDI_ℓ in the datasets (without Vietnamese). The first two metrics (rows) have high correlation and are also statistically significant.

Metrics	G20	Olympics
In-Degree	0.42 ($p < 0.01$)	0.42 ($p < 0.01$)
Eigenvector	0.43 ($p < 0.01$)	0.46 ($p < 0.001$)
Out-Degree	0.96 ($p < 0.5$)	0.29 ($p < 0.03$)
Betweenness	0.07 ($p < 0.6$)	0.08 ($p < 0.55$)
Closeness	0.29 ($p < 0.04$)	0.20 ($p < 0.14$)
Weighted In-Degree	0.09 ($p < 0.55$)	0.08 ($p < 0.57$)
Weighted Out-Degree	0.09 ($p < 0.5$)	-0.19 ($p < 0.14$)

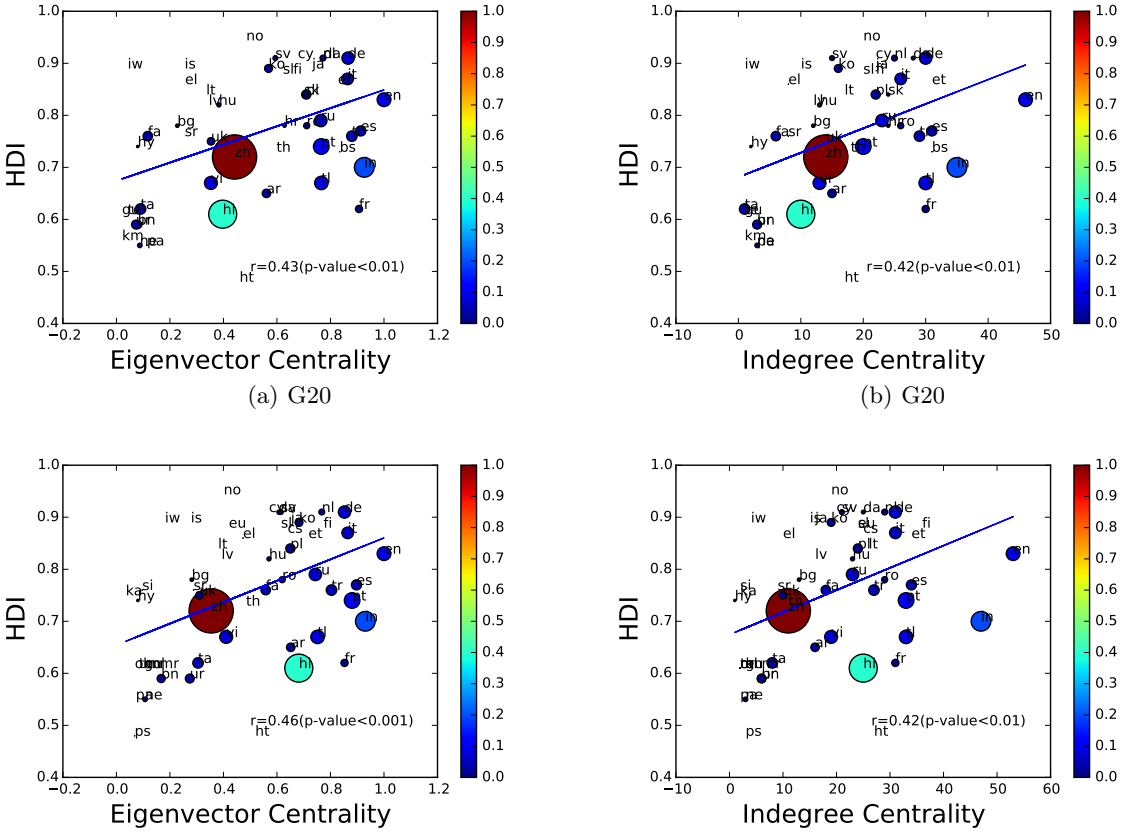
Conclusion and Future Work

In this paper, we set to understand if we can relate the language connection patterns of users on Twitter to standard of living aspects in the real-world, such as HDI. Although it is interesting to extract and analyze the entire Twitter language network, it is rather a very time consuming and expensive procedure. Hence, we used different sets of data from different time periods which demonstrate that the language networks we generated are robust. We also observed that a few popular languages that have very high in-degree and eigenvector centrality do not tend to be the ones with very high HDI. It is worth noting that some of the popular languages are spoken in different parts of the world. We demonstrate that overall the positions of languages correlate significantly with the HDI_ℓ (HDI of the languages). Our work can be extended to understand the other factors that can be added along with the language positions to better describe the variability of the QoL. We also aim to analyze geo-tagged tweets to understand the current location of a user and the relation to the language he chooses to use.

References

Agency, C. I. 2016. The world factbook. `https://www.cia.gov/library/publications/the-world-factbook/fields/2098.html`.

Demirbas, M.; Bayir, M. A.; Akcora, C. G.; Yilmaz, Y. S.; and Ferhatosmanoglu, H. 2010. Crowd-sourced



(a) G20 (b) G20

(c) Olympics (d) Olympics

Figure 6: The significant positive correlation between language centralities and the Human Development Index indicate that the language positions in the network can be co-related to the QoL. Size and color of the languages represent the standard deviation of the users of the languages in different countries.

sensing and collaboration using twitter. In *World of Wireless Mobile and Multimedia Networks (WoWMoM), 2010 IEEE International Symposium on a*, 1–9. IEEE.

Grimes, B. F.; Pittman, R. S.; Grimes, J. E.; et al. 1988. *Ethnologue: Languages of the world*, volume 13. Summer Institute of linguistics Dallas (TX).

Kulshrestha, J.; Kooti, F.; Nikravesh, A.; and Gummadi, P. K. 2012. Geographic dissection of the twitter network. In *ICWSM*.

McLaren, L. 2007. Socioeconomic status and obesity. *Epidemiologic reviews* 29(1):29–48.

Milenkovic, N.; Vukmirovic, J.; Bulajic, M.; and Radojicic, Z. 2014. A multivariate approach in measuring socio-economic development of MENA countries. *Economic Modelling* 38:604–608.

Mocanu, D.; Baronchelli, A.; Perra, N.; Gonçalves, B.; Zhang, Q.; and Vespignani, A. 2013. The twitter of babel: Mapping world languages through microblogging platforms. *PloS one* 8(4):e61981.

Rebello, L. 2015. Un human development index report: Norway leads for the 12th year; uk comes in 14th.

Ronen, S.; Gonçalves, B.; Hu, K. Z.; Vespignani, A.; Pinker, S.; and Hidalgo, C. A. 2014. Links that speak: The global language network and its association with global fame. *Proceedings of the National Academy of Sciences* 111(52):E5616–E5622.

Saha, P., and Menezes, R. 2016a. Exploring the world languages in twitter. In *2016 IEEE/WIC/ACM International Conference on Web Intelligence*, number DOI 10.1109/WI.2016.30, 153–160. IEEE Computer Society.

Saha, P., and Menezes, R. 2016b. A language-centric study of twitter connectivity. In *Social Informatics*, volume 10047. Springer International Publishing. 485–499.

Sandrin, A. 2011. Comparing the hdi with gdp. `http://www.tesionline.it/consult/brano.jsp?id=11583`.