Global Influence of Languages

Zarana Parekh* and Mahima Achhpal†
Information and Communication Technology
Dhirubhai Ambani Institute of Information and Communication Technology
Gandhinagar, India
Email: *201301177@daiict.ac.in, †201301199@daiict.ac.in

***Abstract*—Determining the likelihood of the spread of an idea originating in a particular language can provide useful insights for driving decisions related to diverse areas. In this paper, we determine the languages which are most important in terms of information exchange and the ones which are likely to be isolated. For this purpose, we have studied and implemented the global language networks (GLNs) of Wikipedia articles, Twitter users and book translations, based on the work of Macro Connections, MIT Media Lab described in [4] [1]. To validate the findings, we study the correlation between determined language influence and number of famous speakers of a language based on their Wikipedia biographies. We find a high correlation between the two. For the possibility of extending our observations to other mediums, we also determine the correlation between different GLN datasets, which is also high, indicating a strong correlation between the influence of languages across different mediums.**

I. Introduction

One of the major factors which diversifies the global population is the language and yet it is able to bring people together through the flow of information. Geography plays the most significant role in determining the language spoken by a given section of the population. Over time, cultural exchanges and migration of people lead to the spread of languages across different regions, resulting in varied levels of influence for each language in the world.

In the modern times, information exchange is indispensable for the existence of all spheres of life such as economy, technology, politics and intellectual growth. Also, it is important to understand the languages which are influential in the world or in a region when one is trying to propagate an idea so that it reaches to the maximum number of people. The idea may be in the form of an article, a tweet or any form of document. By studying the global network of bilingually influential people, it is possible to determine the languages which are most influential around the world. Hence, studying the global influence of languages is of utmost importance.

II. Global Influence of Langauges

The spread of an idea depends on the number of speakers of that language and the how well that language connects with other languages. Languages which are co-spoken by a majority of the population considered, are more likely to act as central hubs in such networks and connect the less influential languages. The degree of connectivity of a language in a global network thus helps to determine the language in which to create content or translate to, from the less connected and hence less influential languages.

For our analysis we have considered three global language networks (GLN) - Twitter, book translations and Wikipedia articles; for two reasons. First, a single GLN cannot be used to capture the variations in the influence of a language globally due to the difference in the nature of communication across these platforms. For instance, a colloquial language would be more preferable for informal communications on platforms such as Twitter but less preferable for technical articles. Thus, we have considered these three networks with varying levels of formalism. Second, a geographically restricted medium such as the Chinese analog of Facebook called renren.com fails to capture the global scenario. One limitation to this approach is that it takes into account mostly the online population, thus ignoring verbal and written communication that takes place offline.

A. *Structure of the network*

For the three networks, each node represents a language. The book translations network has directed edges with each edge weighted by the number of translations from the target to the source language. In the undirected Wikipedia network, the end-points of an edge represent the languages in which an author has edited any article, weighted by the number of such authors. Similarly, in the Twitter network, each undirected edge is weighted by the number of people who have tweeted in both languages, represented by the edge end-points.

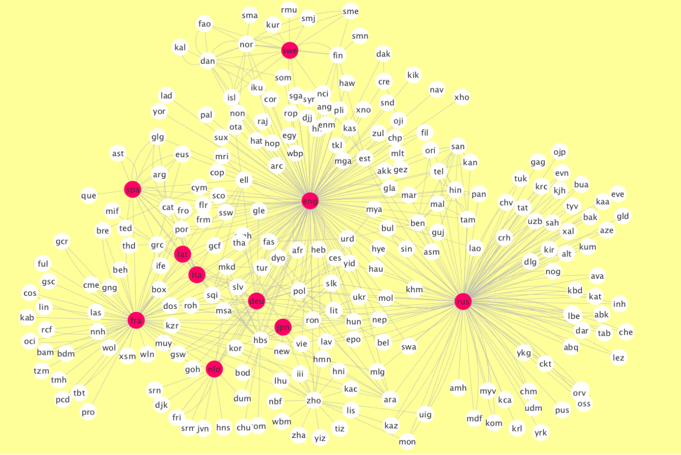

Fig. 1. Visualizing the books translation dataset. Red nodes represent the central hubs.

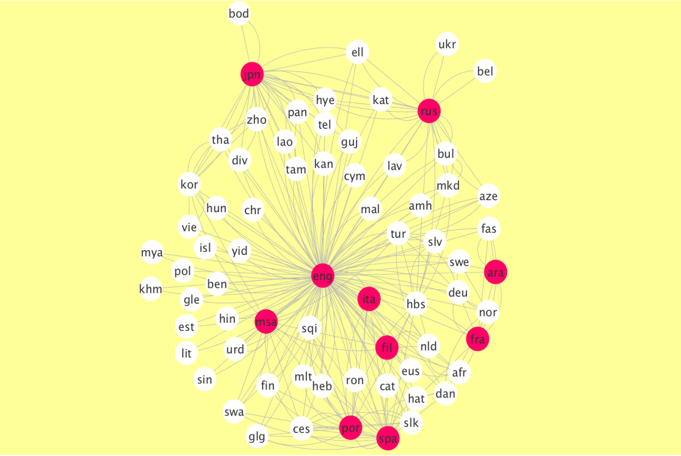

Fig. 2. Visualizing the Twitter dataset. Red nodes represent the central hubs.

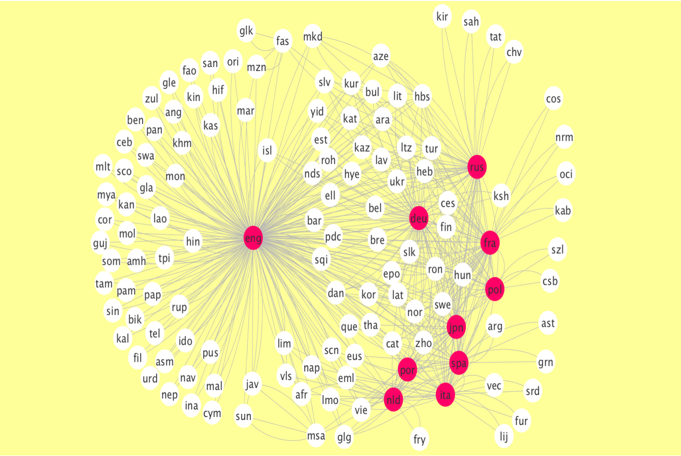

Fig. 3. Visualizing the Wikipedia dataset. Red nodes represent the central hubs.

In all the 3 GLNs we see emergence of central and intermediate hubs which connect the GLN and are critical for information exchange. They are not only vital for information exchange but also are necessary for connecting several other languages in the medium, some of which only have connections to these central hubs. A majority of languages do not have significant links to more than one language in the network. This could be due to the localized nature of their influence.

B. *Determining the most influential languages*

We determine the influence of a language based on how well it is connected to other languages in terms of its in-degree, that is, the number of languages to which an idea originating in this language propagates directly. Also, it is important to consider the influence of the neighbouring languages for the idea to spread globally. Thus, we use the eigenvector centrality as it takes into account the above-mentioned factors while assigning importance to any node. Betweenness centrality has not been used because it assigns equal importance to all neighbours of a node irrespective of their importance in the network. Hence, it fails to consider the characteristics of the neighbours. Also, the range of the betweenness centrality measure is smaller which makes it difficult to differentiate between nodes of similar importance. We have used the in-built algorithm of NetworkX library to determine the eigenvector centrality values. [2]

TABLE I
TOP 10 INFLUENTIAL LANGUAGES OF ALL 3 GLNS DETERMINED BY EIGENVECTOR CENTRALITY (EV)

Rank	Translations	EV	Wikipedia	EV	Twitter	EV
1	English	0.897	English	0.659	English	0.688
2	French	0.297	German	0.478	Malay	0.487
3	German	0.263	French	0.337	Spanish	0.357
4	Italian	0.093	Spanish	0.287	Portuguese	0.346
5	Russian	0.086	Italian	0.157	Filipino	0.137
6	Spanish	0.085	Russian	0.151	Dutch	0.106
7	Japanese	0.043	Dutch	0.134	Arabic	0.056
8	Dutch	0.039	Japanese	0.123	Japanese	0.044
9	Latin	0.034	Portuguese	0.109	French	0.038
10	Swedish	0.033	Polish	0.093	Italian	0.026

Table-1 describes the most influential languages in each medium on the basis of eigenvector centrality. From the table, it can be seen that some of the languages are very influential irrespective of the medium considered. This can be accounted to other economic and historic reasons. For example, due to colonization by the British Empire, English became immensely popular globally.

C. *Determining the isolated languages*

An isolated language is one through which the further propagation of ideas is limited. Hence, the number of translations into and from this language or the cooccurrence values are likely to be lower. Hence, while building the network links which had relatively lower weight and thus were pointed to nodes which were likely to be less influential and isolated, have not been considered in the cleaned data set. Also, nodes with the least eigenvector centralities can also be considered to be isolated.

For the undirected GLNs, we can determine the languages which can possibly be isolated based on their eigenvector centrality values because it captures the interaction between the 2 languages in that network. But to find isolated languages in the book translation network, we have converted it to an undirected network with weights equal to the sum of the interactions (translations in both directions) and then determined the corresponding eigenvector centrality.

There are a large number of languages which would fall into the isolated category due to their insignificant links in the network or lower eigenvector centralities, as can be observed from Fig. 1-3. A list of the same has not been included due to its exhaustive nature.

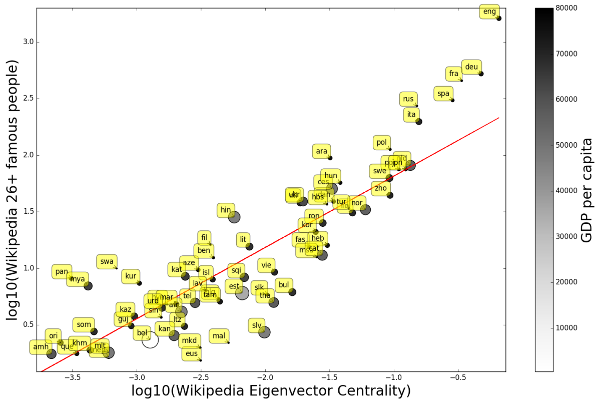

Fig. 4. Wikipedia GLN: Correlation between the position of a language and the influence of its speakers

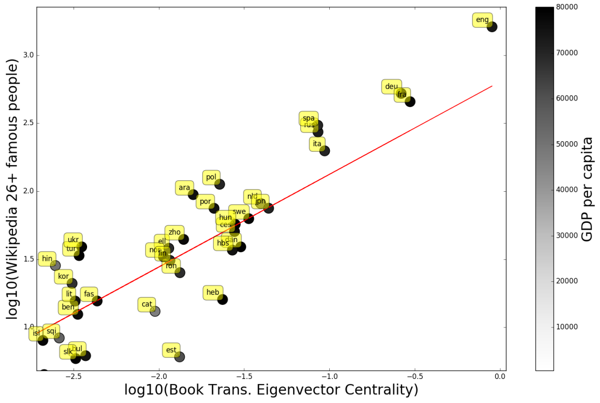

Fig. 6. Book Translations GLN: Correlation between the position of a language and the influence of its speakers

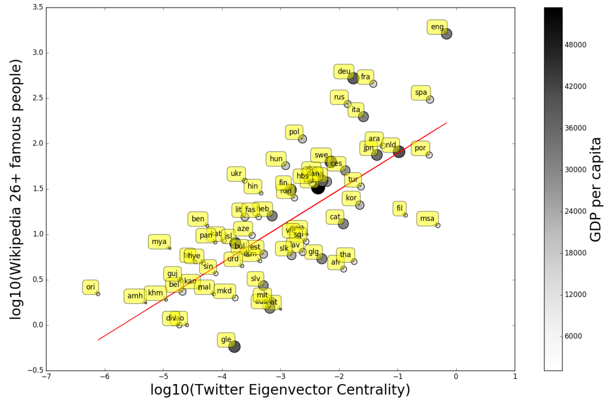

Fig. 5. Twitter GLN: Correlation between the position of a language and the influence of its speakers

D. *Correlation between famous people and influence of languages*

People with biographies on Wikipedia in 26 or more languages are considered to be famous. Detailed explanation for obtaining the dataset can be found in [3]. On the basis of their place of birth, the corresponding language was considered to have contributed to the spread of his or her ideas. Thus, we find the correlation between the influential languages obtained from the GLNs and the number of famous people in each language. Here, the contribution of each person is considered with respect to the language demographics of that nation. For instance, if a person is born in a nation where two languages are spoken in the proportion of 0.63 and 0.39 respectively, then the contribution of that person will be added proportionately to both the languages.

TABLE II
CORRELATION COEFFICIENT BETWEEN NUMBER OF FAMOUS PEOPLE AND THE INFLUENCE OF A LANGUAGE

GLN	Correlation
Book Translations	0.863879620437
Wikipedia	0.865011517954
Twitter	0.718697013766

The high positive value of Pearson's correlation coefficient obtained for the 3 GLNs is given in **Table-2** indicates a direct relationship between the number of famous people and the influence of a language. Such a behaviour is also expected since ideas written in an influential language are likely to spread wider, thus resulting in a greater number of famous people. Similarly, a person is likely to be more famous if his/her work reaches larger number of people, that is it uses a more influential language.

Fig 4-6 represent the correlation between the position of a language in the GLNs and the influence of its speakers which support our observation of high correlation values. GDP per capita which is represented by the bubble size in these figures do not show any clear trend on moving from less to more influential languages. Such a behavior can be accorded to the presence of several factors that contribute to the influence of any language and hence GDP alone is not sufficient to predict the influence of any language.

E. *Similarity between the GLNs*

The high correlation coefficient indicates that the general behaviour of a language in any GLN would be similar. We expect similar results for different mediums for information exchange. Thus, the influence of a language is not biased to the medium. **Fig. 7-9** supports the afore-mentioned notion.

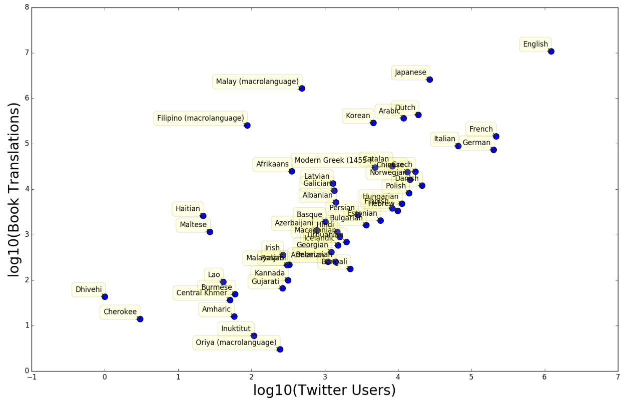

Fig. 7. Similarity between the Twitter and Book Translations datasets

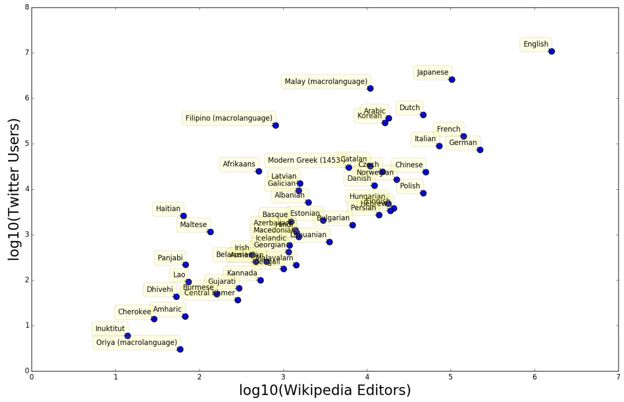

Fig. 9. Similarity between the Twitter and Wikipedia datasets

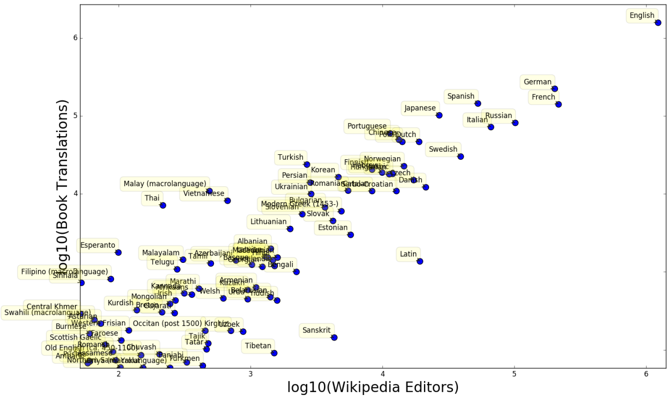

Fig. 8. Similarity between the Wikipedia and Book Translations datasets

III. Conclusion

We begin with creating Global Language Networks (GLNs) representing the co-expressions of languages across the three different mediums of Wikipedia, Twitter and Book translations. The nodes in the network represent different languages and we consider number of translations of articles between pairs of languages in the Wikipedia network, number of users who have co-tweeted in a pair of languages for the Twitter GLN and number of pairwise translations of a book from its source language to other languages (hence, directed) for the book translations GLN, as the weights of the edges.

By visualizing these GLNs in **Fig. 1-3**, we observe emergence of central and intermediate hubs representing most influential languages in the network. Then eigenvector centrality is used as the measure for determining the influence of each language in the network as it considers not only the importance of a language, but also of its direct and indirect neighbours iteratively. Lower eigenvector centralities and less significant connections are used to identify the possible language isolations in the networks.

To validate our findings, we determine the correlation between the influential languages and the number of their famous speakers based on the Wikipedia biographies [3]. **Fig. 4-6** and **Table 2** show high correlation between the two and hence support our observations of influential languages. The observed similarity between the GLN datasets in **Fig. 7-9** serve as another measure of validation that our results can be extended to other mediums as well.

For further exploring the nature of the network, temporal changes can be considered which can even help to make more accurate observations and predictions about language isolations and evolutions in the network.

The source code and data used can be found **here**.

References

[1] Global language network. http://language.media.mit.edu/visualizations/books. Accessed: 2016-11-06.

[2] Networkx: Documentation on eigenvector centrality. https://goo.gl/3JeYGY. Accessed:2016-11-06.

[3] Supplementary reading for the paper on global language network. https://goo.gl/Myqv4f. Accessed: 2016-11-06.

[4] Shahar Ronen, Bruno Gonçalves, Kevin Z Hu, Alessandro Vespignani, Steven Pinker, and César A Hidalgo. Links that speak: The global language network and its association with global fame. *Proceedings of the National Academy of Sciences*, 111(52):E5616–E5622, 2014.