Language Relations on Twitter: A Network Science Approach

by

Priya Saha

A dissertation submitted to Florida Institute of Technology in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science

Melbourne, Florida April 2017
We the undersigned committee hereby recommend that the attached dissertation be accepted as fulfilling in part the requirements for the degree of Doctor of Philosophy in Computer Science

Language Relations on Twitter: A Network Science Approach

by

Priya Saha

Ronaldo Menezes, Ph.D.
Program Chair, Computer Sciences
Advisor and Committee Chair

Ryan Stansifer, Ph.D.
Associate Professor, Computer Science

Pat Bond, Ph.D.
Associate Professor, Computer Science

Lisa Steelman, Ph.D.
Associate Professor, Department of Psychology
Online social networks (e.g., Twitter and Facebook) play a vital role in the spreading of information in today’s world. Interestingly, the spread of information is enabled by the existence of an underlying connectivity of the users. One factor influencing the online connectivity, which only recently has been receiving attention, is the language used by the user in his or her activities. The understanding of information propagation from the perspective of languages is of particular interest because we live in a world with a very diverse set of languages. Using Network Science approaches, we demonstrate that Twitter users have a strong preference to connect to people who use their own language, but more importantly, we found that this preference is stronger than the tendency to connect to people with a similar popularity level (i.e., the traditional notion of homophily). The connecting patterns between users of different languages vary considerably and such patterns shed light on the similarity of languages from a user-preference point of view. Furthermore, we unveil the “Twitter Language Network”, a connected system of many different languages, and we analyze several of its interesting characteristics. In addition, we demonstrate that the position of languages in the Twitter Language Network correlates with the social and development indicator of the language users. This dissertation presents an investigation of the language structure of Twitter, giving a better understanding of the connectivity of users in the context of their languages.
# Contents

1 Introduction .................................................. 1

2 Literature Review ............................................. 3
   2.1 History of Networks ........................................ 3
   2.2 Related Metrics ........................................... 5
      2.2.1 Path .................................................. 5
      2.2.2 Degree Centrality .................................... 6
      2.2.3 Eigenvector Centrality ............................... 6
      2.2.4 Betweenness Centrality ............................... 7
      2.2.5 Closeness Centrality ................................ 8
      2.2.6 Clustering Coefficient ............................... 8
      2.2.7 Degree Distribution ................................. 8
      2.2.8 Assortativity ........................................ 9
      2.2.9 Entropy .............................................. 10
      2.2.10 Small-World Network ............................... 10
   2.3 Related Works ............................................. 11
   2.4 Literature Gap ............................................ 20
      2.4.1 Problem Statement ................................... 21
      2.4.2 Thesis Statement ..................................... 22

3 Datasets and Network Validation ............................... 23
   3.1 Datasets .................................................. 23
3.2  Network Generation ................................................. 24
   3.2.1  Dataset 1: Follower Network .................................. 24
   3.2.2  Dataset 2: Retweet Network ................................... 26
3.3  Network Data Validation ......................................... 32

4  Language-Centric User Connectivity Analysis .......................... 36
   4.1  Assortativity of Users ........................................... 38
       4.1.1  Degree Assortativity ..................................... 38
       4.1.2  Language Assortativity .................................. 40
   4.2  Diversity of Users .............................................. 41

5  Exploring the World Languages ...................................... 46
   5.1  Reachability of Languages ...................................... 47
   5.2  Similarity of Languages ........................................ 50
       5.2.1  Similarity Based on the Existence of Connections .... 50
       5.2.2  Similarity Based on the Weight of Connections ....... 52

6  Language as a Measure of Welfare ................................... 58
   6.1  Tweet Collection and Network Generation ....................... 59
       6.1.1  Additional Dataset: Olympic Games 2016 ............... 59
   6.2  HDI of Languages ............................................. 65
   6.3  Language Rankings by Users and HDI .......................... 66
       6.3.1  Languages by the Number of Users ..................... 66
       6.3.2  Languages by HDI ........................................ 66
   6.4  Correlation between Network Properties and HDI ............. 68

7  Conclusions and Future Work ........................................ 71

A  List of Publications ................................................ 74
List of Figures

2.1 A simple example of an undirected network ................................................. 4

2.2 A directed network. The edge labels indicate the edge weight. The edges with
higher weight are thicker ................................................................................. 5

2.3 An example of a degree distribution with power-law ..................................... 9

2.4 Average clustering coefficient of users (source: Myers et al. [1]). Mutual degree is
the degree of the mutual graphs ...................................................................... 13

2.5 Influence of users across the measures (source: Cha et al. [2]). ..................... 13

2.6 Characterization of users based on their follow counts (source: [3]). .............. 14

2.7 Distribution of users of the top ten countries on Twitter (source: Poblete et al. [4]). 17

2.8 Distribution of top three languages in the top ten countries on
Twitter (source: Poblete et al. [4]). ................................................................. 17

2.9 Distribution of French and English users in
Montreal, Canada (source: Mocanu et al. [5]) .................................................. 19

3.1 A sample network of Vladimir Putin. Nodes represent users and edges represent
the follower relation. The size of the nodes represents the degree of the users. In
the network above, we represent the followers as well as the followers of followers
of Vladimir Putin ......................................................................................... 25

3.2 A small sample follower network of a few famous Russian (nodes of green color)
egos. The color of nodes represents languages and the size of nodes represents
degree ............................................................................................................. 27
3.3 FLN. The color of a node represents the betweenness of the node in the network: color ranges from green (low betweenness) to red (high betweenness). The size of a node represents degree of the node. The width of an edge represents the weight of the edge. In the network above, most of the information passes through English (en), followed by Spanish (es), Italian (it), Russian (ru), German (de), Portuguese (pt), French (fr), Swedish (sv), Turkish (tr), and Indonesian (id). The aforementioned languages are also very strongly connected in the network.

3.4 Snapshot of the tweets that contain the last names of leaders of the G20, that are posted by Twitter users.

3.5 RLN. The color of a node represents its betweenness in the network: color ranges from green (low betweenness) to red (high betweenness). The size of a node represents the degree of the node. The width of an edge represents the weight of the edge. In the network above, most of the information passes through English (en), followed by Spanish (es), Italian (it), Russian (ru), German (de), Portuguese (pt), French (fr), Swedish (sv), Turkish (tr), Arabic (ar), Chinese (zh) and Dutch (nl). The aforementioned languages are also very strongly connected in the network.

3.6 Comparison of the distributions demonstrating that the log-normal distribution best describes the degree distribution of the follower network. Table 3.5 shows the comparison results.

3.7 Comparison of the distributions demonstrating that the log-normal distribution best describes the degree distribution of the retweet network. Table 3.6 shows the comparison results.

4.1 The top 20 languages according to the percentage of Twitter users in the two networks studied; that is, the popularity of the languages is represented here.

4.2 Illustration that the nodes with a low degree have a high average neighbor degree and vice versa in the follower network, indicating disassortativity.

4.3 Illustration that the nodes with low degree have high average neighbor degree and vice versa in the retweet network, indicating disassortativity.
4.4 The darker part of the column represents the total number of edges that connect users who use the same language (indicated in the x-axis). The lighter part of the column represents the total number of edges in the network that connect the users who use other languages (different from the one in the x-axis).

4.5 A toy example to demonstrate the language exposure vector.

4.6 The entropy of the languages in the networks.

5.1 Shortest paths between languages in the FLN. The diameter of the FLN is 4.

5.2 Shortest paths between languages in the RLN. The diameter of the RLN is 3.

5.3 FLN: the size of a node is based on its degree and the color of the node represents the block it belongs to. The nodes in the same block display similar pattern of connections. The tree representation in the FLN shows that there are four subtrees at the first cutting level. The subtrees show the languages that belong to each block.

5.4 RLN: the size of a node is based on its degree and the color of the node represents the block it belongs to. The nodes in the same block have similar patterns of connections. The tree structure in the RLN shows that there are three subtrees at the first cutting level; that is, the important languages including English form a block and the other languages are arranged in two other blocks. The subtrees show the languages of every block.

5.5 Jaccard similarities of the languages in the FLN.

5.6 Jaccard similarities of the languages in the RLN.

5.7 Correlation matrix of languages in the FLN.

5.8 Correlation matrix of languages in the RLN.

6.1 The term “Olympics” was translated to several languages and used to collect tweets related to the Summer Olympic games in Rio de Janeiro, Brazil.
6.2 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 edge width represents the normalized tweet frequency between the languages. The represented language families are: Turkic, Indo-European, Japonic, Tai-Kadai, Austronesian, Uralic, Koreanic, Austrasiatic, Sino-Tibetan, Afroasiatic, Dravidian.

6.3 Language network of the Olympics dataset: the color of a node represents the language family and the size of the node represents the in-degree. The edge width represents the normalized tweet frequency between the languages. The represented language families are: Turkic, Indo-European, Japonic, Tai-Kadai, Austronesian, Uralic, Koreanic, Austrasiatic, Sino-Tibetan, Afroasiatic, Independent, Kartvelian, Dravidian.

6.4 Distribution of users per language in the G20 dataset.

6.5 Distribution of users per language in the Summer Olympics dataset.

6.6 The significant positive correlation between language centralities in the G20 dataset 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.

6.7 The significant positive correlation between language centralities in the Olympics dataset 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.
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Comparison of Twitter graphs with Facebook and MSN (source: Myers et al. [1]). While the users in the Follow Graph may or may not have reciprocal connections, the users in the Mutual Graph must have the reciprocal connections in the network.</td>
</tr>
<tr>
<td>2.2</td>
<td>Top 10 languages on Twitter according to the number of tweets (source: Hong et al. [6]).</td>
</tr>
<tr>
<td>2.3</td>
<td>Percentage of characteristics on Twitter. (source: Hong et al. [6]).</td>
</tr>
<tr>
<td>3.1</td>
<td>List of egos per language in the Follower Network dataset.</td>
</tr>
<tr>
<td>3.2</td>
<td>List of last names of the G20 leaders in the Retweet Network dataset.</td>
</tr>
<tr>
<td>3.3</td>
<td>Description of the two datasets used in this work. A third dataset is used in Chapter 6 only and its description is included in that chapter.</td>
</tr>
<tr>
<td>3.4</td>
<td>Description of the networks that exist in our society and have been studied previously. They differ widely in their sizes. ( N ) refers to the number of nodes, ( L ) refers to the number of edges and (&lt;k&gt;) refers to the mean degree of the networks.</td>
</tr>
<tr>
<td>3.5</td>
<td>The log-likelihood comparative study shows that log-normal best describes the in-degree and out-degree distributions of the follower network.</td>
</tr>
<tr>
<td>3.6</td>
<td>The log-likelihood comparative study shows that log-normal best describes the in-degree and out-degree distributions of the retweet network.</td>
</tr>
<tr>
<td>4.1</td>
<td>The table below shows that the networks are disassortative by degree and assortative by language.</td>
</tr>
</tbody>
</table>
5.1 Top ten languages by degree in the FLN and the RLN. The FLN has total 57 languages while the RLN has 51 languages represented.

6.1 Descriptive statistics about the datasets. Monolingual refers to users who use one language and multilingual refers to users who use more than one language.

6.2 A toy example to demonstrate the tweet percentages of the users in different languages.

6.3 We used the data in Table 6.2 to generate the matrix of users who demonstrate English as their frequent language. From the user matrix, we generated the normalized English frequency, which is the sum of the tweet frequency in each language divided by the number of users.

6.4 We used the data in Table 6.2 to generate the matrix of users who demonstrate German as their frequent language. From the user matrix, we generated the normalized German frequency, which is the sum of the tweet frequency in each language divided by the number of users.

6.5 We used the data in Table 6.2 to generate the matrix of users who demonstrate Swahili as their frequent language. From the user matrix, we generated the normalized Swahili frequency, which is the sum of the tweet frequency in each language divided by the number of users.

6.6 Top languages in the G20 dataset according to in-degree and eigenvector centrality.

6.7 Top languages in the Olympics dataset according to in-degree and eigenvector centrality.

6.8 The five countries with highest and lowest HDI. 

6.9 The five languages with highest and lowest HDI. 

6.10 The correlation of the network metrics with HDI in the datasets. The first two metrics (rows) have high correlation and are also statistically significant.
Declaration

I declare that the work presented in this dissertation is my own work except where cited to another author. The material in this thesis is previously published by the author. For a complete list of publications, please refer to Appendix A at the end of this dissertation.
Chapter 1

Introduction

Many systems exist in the form of connected structures; examples include society (a connected structure of people), ecosystems (a connected structure of who-eats-whom), Internet (a connected structure of computers), and the World Wide Web (a connected structure of webpages), to name a few. A connected structure can generally be referred to as a network; a collection of objects (referred to as nodes) linked by their interactions (referred to as edges). A network of people formed by their social interactions is called a social network. These networks can provide insights on how people gather news, form opinions, and other interesting phenomena, such as the spread of disease. In recent years, we have observed the rapid emergence of online social networks, contributing to information travelling much faster and wider than just a few years ago. For instance, Facebook was heavily used during the Arab Spring in 2010 [7] and Twitter was crucial during the U.S. presidential elections in 2008 [8]. Twitter has even been shown to work as a real-time sensor to events such as earthquakes [9].

Though the diffusion of information in online social networks is a popular area of interest, the research community has not paid much attention to some hindrances that could be imposed to curtail information flow. Language is a crucial medium for the global spread of information because users generally need to understand the subject of the information before passing it forward. Questions such as the following are yet to be answered: “Which languages are influential in the transmission of information?”, “Does the position of a language in a network play a role in its exposure to information?”, “Are there different patterns of connectivity for users of different
languages?”, “Are users organized in social networks according to their languages?”, and “Is the pattern of language connectivity subject-dependent?”.

In this context, understanding the structure of languages in online social networks that emerges from the dynamics of user connectivity is very important as it can be an enabler for information spread. Using concepts from Network Science and Information Theory, this dissertation focuses on understanding the connectivity of Twitter users as a function of their languages. The increase in the number of registered users in the past few years indicates the growing popularity of Twitter as a social network service [10]. We chose to explore Twitter not only because it has become a major source of information to many people but also because it has gained immense interest among the research scientists who are trying to understand specific aspects of human behavior. Despite known sampling bias, Twitter has become a testbed for confirming or finding new theories related to human behavior [11]. Furthermore, the diversity of languages that Twitter supports is an interesting source for understanding the language relations. Since Twitter has a mix of users of many different languages, it is reasonable to assume that a highly connected language structure (network) exists underlying the social connectivity of the users. This language structure is formed by the multilingual users who act as bridges between languages (e.g., they may read information in one language and post in another). Unlike the social structure of friends, the structure of languages of friends is harder to be extracted.

We organize our study and findings as follows: In Chapter 2, after a review of the literature, we describe the metrics that we used in our analysis; we also discuss the research gap followed by the thesis statement. Next, we describe the methodology used for collecting data in Chapter 3; we also demonstrate how we generated and validated our networks given that our datasets are smaller than the real Twitter network. Chapter 4 shows the dynamics of the users and how they tend to associate with others who are like them from the language perspective. Chapter 5 explores the world languages; the analysis includes understanding of similarities of languages in the context of the existence of as well as the weight of the connections between them. Chapter 6 demonstrates how language networks can be used as a proxy to understand the development of societies. Finally in Chapter 7, we discuss the contributions and limitations of this work as well as suggestions for future work.
Chapter 2

Literature Review

2.1 History of Networks

The field of Network Science borrows many of its concepts and definitions from graph theory. Graph theory has its beginnings in 1736 with Euler and his famous solution to the Seven Bridges of Königsberg problem. Euler used a mathematical representation based on nodes and edges to demonstrate that there was no solution to the problem of finding a path around the city of Königsberg using all its bridges without repeating one [12]. Since then, the study of graphs has become popular due to its ability to model real-world problems related to many branches of science. Network science can be defined as the study of the interconnections of objects in the real world in the form of a network and the dynamics of this connected structure. Networks can be directed (edges represent a directional relation between the nodes) or undirected (edges represent a reciprocal relation between the nodes). Figure 2.1 represents an undirected network and Figure 2.2 represents a weighted directed network. The weight of an edge in a weighted network indicates the number of interactions between the objects. One of the most common uses of networks as a framework for modeling connections is the study of social interactions. In 1930, Moreno became actively involved in the study of social dynamics among people [13]. His studies and results are one of the first published works in what is now called social network analysis. Social networks refer to interaction among a group of people, but over time, the concept of social networks became a synonym for online social networks, which are considered as today’s
main media for writing and exchanging ideas between people. Some of the popular online social networking sites are Facebook, Twitter, and LinkedIn. People form communities and share content according to their own interests. There are many works that demonstrate that most online social networks have similar characteristics [14–16]. In Section 2.2 we describe several metrics that help in our understanding of the structure of networks and social interactions [17,18].

![Diagram of an undirected network](image)

Figure 2.1: A simple example of an undirected network
2.2 Related Metrics

2.2.1 Path

A path is defined as a sequence of nodes such that every consecutive node in the sequence is connected by an edge. In a directed network, each edge in the path must be traversed respecting the direction of the edges. In an undirected network, the edges can be traversed in either direction.

The length of a path in a network is the total number of edges in the path. The shortest path between two nodes in a network is called the \textit{geodesic path}. The longest of all geodesic paths represents the \textit{diameter} of the network. If there is no path between two nodes in a network, then the distance between the nodes is considered infinite. In Figure 2.1 and Figure 2.2, $A \rightarrow B \rightarrow E \rightarrow F$ represents a path.
2.2.2 Degree Centrality

The degree of a node is defined as the number of edges (connections) incident upon it in the network. For example, in Figure 2.1 both B and C have degree 5, which is the highest degree. In a directed network, degree is analyzed in two ways: in-degree and out-degree. In-degree is the number of edges incoming to a node whereas out-degree is the number of edges outgoing from the node. In Figure 2.2 B has in-degree 1 and out-degree 4. Additionally, in a weighted directed network, we can also compute the weighted in-degree and weighted out-degree. Weighted in-degree of a node is the sum of the weights of the incoming edges of the node and weighted out-degree is the sum of the weights of the outgoing edges of the node. In Figure 2.2, the weighted in-degree of B is 2 and weighted out-degree of B is 12. The degree of a node may be used to indicate the popularity of the node in the network (awarding the node one point for each neighbor). In a social network, individuals who have a high degree might have more importance than the individuals who have a low degree. Furthermore, people with a higher degree in a social network may have more access to information circulating in the network. The ratio between the number of edges present and the total number of possible edges in a network is called the density of the network.

2.2.3 Eigenvector Centrality

Eigenvector centrality is a simple extension of the degree centrality. In many cases, the importance of a node is increased by having connections to other nodes that are themselves important, which the degree centrality fails to capture. Eigenvector centrality gives every node a score proportional to the sum of the scores of its neighbors and measures the influence of the node due to many connections, or important connections, or due to both. To calculate the eigenvector centrality of a node \(i\) in an unweighted network, we can start by setting the centrality \(x_i = 1\) for all \(i\). Although \(x_i\) is not a useful measure, it can be used to compute another metric \(x_i'\), which can be defined as the sum of the centralities of \(i\)'s neighbors,

\[
x'_i = \sum_j A_{ij} x_j,
\]

(2.1)
where, \( A_{ij} \) is an element of the adjacency matrix \( A \). In a weighted network, \( A \) represents the weighted adjacency matrix, in which each element represents the weight of the edges. Note that eigenvector centrality is an iterative process and not the weight of the edges only. If we use the linear combination of eigenvectors, the above equation can be represented as

\[ x'_i = k_1^{-1} \sum_j A_{ij} x_j, \]  

where, \( k_1 \) is the largest eigenvalue of \( A \). Although eigenvector centrality is an efficient metric, in many cases it might not be the best one. For example, in Figure 2.2 node A only has outgoing edges. Hence, the eigenvector centrality of A is zero. B has an incoming edge and two outgoing edges. The incoming edge of B originates from A. Hence the eigenvector centrality of B is also zero. In this perspective, a node can be pointed to by other nodes, which can be pointed to by many more, and so on, but if the progression ends in a node with in-degree zero, the final value of the centrality will be zero. The problem with the Eigenvector centrality is addressed by Katz Centrality (see [19] for more information on Katz Centrality).

### 2.2.4 Betweenness Centrality

The betweenness of a node is the number of times it appears in the shortest path of two other nodes. It is defined as

\[ B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}, \]  

where, \( \sigma_{st} \) is the total number of shortest paths from node \( s \) to \( t \) and \( \sigma_{st}(v) \) is the total number of shortest paths from node \( s \) to \( t \) passing through \( v \). In a social network, information can be passed from one user to another. Most of the information tends to pass through the users with high betweenness. Hence, the users with high betweenness hold an advantageous position in the network due to their ability to control information passing between users. In Figure 2.1, B has the highest betweenness centrality.
2.2.5 Closeness Centrality

Closeness centrality of a node measures the mean distance of a node to other nodes in the network. It can be given by

\[ y = \frac{n}{\sum_j d_{ij}}, \] (2.4)

where, \( d_{ij} \) is the length of the geodesic path from \( i \) to \( j \) in the network with \( n \) number of nodes. In a social network, a user with low closeness centrality might find that his or her opinions reach other users in the community faster than the rest.

2.2.6 Clustering Coefficient

Let us have a node \( i \) in a network that has \( k_i \) edges (or \( k_i \) immediate neighbors). If the immediate neighbors of \( i \) were part of a clique (a complete subgraph), then there would have been \( k_i(k_i - 1)/2 \) number of edges among the neighbors. The clustering coefficient \( C_i \) of \( i \) is given by the ratio between the number of edges \( (E_i) \) that actually exist between the \( k_i \) nodes and the total number of \( k_i(k_i - 1)/2 \) edges:

\[ C_i = \frac{2E_i}{k_i(k_i - 1)}. \] (2.5)

The clustering coefficient of a network can be measured as

\[ \bar{C} = \frac{1}{n} \sum_{i=1}^{n} C_i, \] (2.6)

where, \( n \) is the total number of nodes. In social networks, the clustering coefficient quantifies the tendency of users to form closed groups of friends. It measures the probability that two neighbors of a user are themselves neighbors. In Figure 2.1, M has the highest clustering coefficient.

2.2.7 Degree Distribution

Not all nodes in a network have the same number of edges. The regularity in the number of edges that nodes have in a network can be represented by a distribution function \( p(k) \). \( p(k) \) gives the probability that a randomly selected node in a network has \( k \) edges. Most real-world networks
exhibit a power-law in their degree distributions \[20,21\]. Networks that have power-law in their degree distributions are called scale-free networks. A network that follows the power-law has a majority of nodes with low degree and a few nodes with high degree (called hubs). Figure 2.3 represents a degree distribution that has power-law. A power-law distribution is given by

\[ p(k) \sim k^{-\alpha} \tag{2.7} \]

where \( \alpha \) is the degree exponent. In real-world networks, \( \alpha \) ranges between 2 and 3 \((2 < \alpha < 3)\). The degree distribution of a network can be represented by a vector \( v = \langle a_i, b_i \rangle, (a_{i+1}, b_{i+2}) \ldots (a_n, b_n) \rangle \) of \( n \) elements, where \( a_i \) and \( b_i \) are the \( i \)-th element of \( v \) and \( a \) represents the number of nodes with degree \( b \). In Figure 2.1, \( v = \langle (2, 5), (1, 4), (2, 3), (3, 2), (6, 1) \rangle \).

2.2.8 Assortativity

Assortativity is a concept in Network Science that describes the tendency of nodes to associate with those that are similar to themselves \[18,22,23\]. A network is said to be assortative if a fraction of edges running between nodes have similar attributes (e.g., age, gender, color). Assortativity can be measured as

\[ r = \frac{1}{\bar{q}^2} \sum_{jk} jk(e_{jk} - q_j q_k), \tag{2.8} \]
given that
\[ q_k = (k + 1)p_{k+1}, \]  
(2.9)
where \( p_k \) is the probability that a randomly chosen node in a network will have degree \( k \), \( \delta_q^2 \) is the variance of the distribution \( q_k \) and \( e_{jk} \) is the joint probability distribution of the remaining degrees of the two vertices at either end of a randomly chosen edge. The value of \( r \) ranges from \(-1 < r < 1\), meaning disassortative to assortative respectively \[24\]. The degree assortativity value of the network in Figure 2.1 is -0.4, indicating the network is disassortative by degree; nodes with high degree tend to connect to other nodes with low degree and vice versa.

2.2.9 Entropy

In Information Theory, entropy measures the disorder or uncertainty in the system \[25\]. Although entropy can be quantified in many different ways, Shannon entropy is one of the most commonly used metrics \[26\]. The Shannon entropy can be measured by

\[
s = -\sum_m p_m \ln(p_m),
\]  
(2.10)
where \( p_m \) is the probability of occurrence of the \( m \)-th possible value of the system. Higher entropy indicates a more disordered (more uncertain) distribution, while a lower entropy indicates the distribution is more ordered (less uncertain).

2.2.10 Small-World Network

In 1967, the famous psychologist Stanley Milgram performed an experiment designed to determine the distance between people in a social network. Milgram chose 96 random participants in Omaha, Nebraska and sent a set of packages to each of the participants. Milgram also chose a target individual in Boston. The Nebraska participants were told to send the package to the participant in Boston. According to the instructions, if any Nebraska participant happened to not know the Boston participant, the former could only forward the package to someone he or she knew on a first-name basis to help the parcel reach the target person. This experiment gave birth to the idea that any two people in the world are separated by just a few intermediary
friends. This discovery led to the concept of small-world networks, which correspond to the social structure seen in large groups of people. A small-world network has a low average path length and a high clustering coefficient [16].

2.3 Related Works

In this thesis, we used the metrics above to characterize the structure of language networks and the social interactions on Twitter. Twitter has several interesting features, such as tweet, retweet, followers, hashtags, mention and replies. A follower is a user who subscribes to receive updates from other users. Twitter users who want to follow other users do not require approval from the latter. A tweet is a message of 140 characters that can be posted on Twitter by a user. A user can share a tweet from another user, called retweet and forward it to his or her own followers. A hashtag is a phrase preceded by a # symbol. A hashtag enables users to follow a topic easily. A mention is a tweet containing the name of a user (@username) in the content. A reply is a response to a tweet posted by another user. While registering for the Twitter service, a user has to declare his or her language preference in the profile, which we refer to as the user language in our study. The assumption here is that if a user declares the choice of language to interact on Twitter, he or she understands the language. Social network scientists have published works that shed light on the spread of information and the role of multilinguals on Twitter [28–30]. Below, we describe a few works that motivated us in our study.

Kwak et al. identified influential users on Twitter by ranking them according to their number of followers and retweets [31]. The ranking of users varied in both the cases, indicating that users who have many followers are not always the ones who post interesting tweets. They constructed a follower network (a network of users where the edges represented the follower relationship) and argued that the follower network exhibited a non-power-law distribution, a short diameter as well as low reciprocity. Their findings demonstrated that popular Twitter users are mostly celebrities and do not tend to follow their users back; in fact, only 22.1% users have reciprocal relations between them; hence Twitter is a source of information to many users rather than just a social networking site.
Myers et al. investigated two types of Twitter follower networks and demonstrated that they exhibit the structural characteristics of both the social and information networks\(^1\). The authors constructed a Twitter Follow Graph, in which the users have a follower relationship; the relations may or may not be reciprocal. In addition, they also constructed a Twitter Mutual Graph, which consisted of users who have a reciprocal follower relationship. Furthermore, they also examined the Mutual Graphs of three different countries—Brazil (BR), Japan (JP), and the United States (US). The reciprocity of the Japan follower network was found to be much higher than the US or Brazil. In Japan, if a user follows another user, it is highly likely that the latter will also get followed back. The authors used a graph metric called spid for their findings. Spid is defined as the dispersion of the path length distance of the network. Social networks have a spid of less than one. Table 2.1 shows that the spid of the Twitter networks is less than one and slightly higher than Facebook. Hence, the Twitter distribution is wider than Facebook. Like other social networks, the local clustering coefficient of the Twitter networks decreased with an increasing degree (Figure 2.4). The authors suggested that when a new user subscribes to Twitter, he or she usually follows the popular Twitter profiles that have high in-degree in order to gain information from the network. Eventually, as the user becomes more “experienced” by following many other users and by getting followed as well, he or she becomes selective of whom to follow and engages with other users who have similar interests. Moreover, a comparative study performed by Lerman et al. also showed that the follower network structure is very similar to other social networks such as Digg\(^32\).

Table 2.1: Comparison of Twitter graphs with Facebook and MSN (source: Myers et al.\(^1\)). While the users in the Follow Graph may or may not have reciprocal connections, the users in the Mutual Graph must have the reciprocal connections in the network.

<table>
<thead>
<tr>
<th>Graph</th>
<th>Avg. Path Length</th>
<th>Spid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Follow Graph</td>
<td>4.05</td>
<td>0.12</td>
</tr>
<tr>
<td>Mutual Graph</td>
<td>4.17</td>
<td>0.11</td>
</tr>
<tr>
<td>Mutual BR</td>
<td>3.78</td>
<td>0.13</td>
</tr>
<tr>
<td>Mutual JP</td>
<td>3.89</td>
<td>0.16</td>
</tr>
<tr>
<td>Mutual US</td>
<td>4.37</td>
<td>0.18</td>
</tr>
<tr>
<td>Other Networks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Facebook</td>
<td>4.74</td>
<td>0.09</td>
</tr>
<tr>
<td>MSN</td>
<td>6.6</td>
<td>-</td>
</tr>
</tbody>
</table>
Figure 2.4: Average clustering coefficient of users (source: Myers et al. [1]). Mutual degree is the degree of the mutual graphs.

Cha et al. analyzed the influence of users in terms of in-degree, number of retweets, and mentions in the Twitter follower network [2]. They used three metrics to understand the influence of users: in-degree influence, which represents the popularity of a user; retweet influence, which represents the content value of the tweet of a user; and mention influence, which measures the name value of a user. The authors discussed three findings in their work. First, there is little overlap among the top users according to retweets, in-degree, and mentions. Figure 2.5 shows that there is much less overlap across the influence metrics of the users because the influence metrics capture different types of importance of the users. Second, the authors show that influential users hold their influence over a wide range of topics. This finding could help understand the effectiveness of employing the influential users for advertisements. Third, ordinary users can gain influence by focusing on a single topic and posting creative tweets rather than simply conversing with other users.

Figure 2.5: Influence of users across the measures (source: Cha et al. [2])
In another study, Cha et al. proposed grouping of users based on their distribution of links in the follower network. According to the authors, users can be categorized into (i) mass media or the extremely well-connected users, (ii) grassroots or the users with no more than 200 users, and, (iii) evangelists or the well-connected small group of users \[33\]. The mass media users do not follow many users but they are followed by many. The grassroot users have a huge presence in the network although they do not have many followers. Evangelists are the users who actively take part in information flow in the network.

Krishnamurthy et al. investigated the Twitter follower network and demonstrated another characterization of users on Twitter, namely: (i) broadcasters or the users who have a very high number of followers, such as the New York Times, the BBC and other media outlets, (ii) acquaintances or the users who have reciprocity in their connections, and (iii) spammers or the users who tend to connect to as many people as they can and hope to have reciprocal relations \[3\]. In Figure 2.6, broadcasters are represented in the extreme left. Although broadcasters usually have a very large number of followers, their relations are not always reciprocal. The acquaintances are represented by the large cluster in the middle of the figure. Miscreants are represented at the extreme right of the figure, where the users show a very high following in comparison to the low number of followers.

Figure 2.6: Characterization of users based on their follow counts (source: \[3\]).
Social networks tend to favor connections among users of similar or dissimilar characteristics, a phenomenon called assortative mixing, as we discussed before. Bollen et al. inspected a Twitter follower network and found that psychological states of users (such as happiness) can be assortative on Twitter\footnote{34}; users who are happy connect to others who also tend to be happy and the same is true for unhappy users. The authors used a metric called subjective well-being (SWB) to understand the happiness level of the users in the Twitter follower network. Below, Subjective well-being $S(u)$ of a user $u$ can be defined as

$$S(u) = \frac{N_p(u) - N_n(u)}{N_p(u) + N_n(u)},$$

(2.11)

where $N_p(u)$ is the number of positive terms in the tweet of the user and $N_n(u)$ is the number of negative terms in the tweet of the user. Using Equation (2.11), the SWB of every user was computed. The authors calculated two types of assortativity: “pairwise node assortativity” and “neighborhood assortativity” of the network. Pairwise node assortativity can assess the degree to which two connected users have similar SWB values and neighborhood assortativity can assess the overall SWB of all the users a particular user interacts with. They found that the pairwise node assortativity of their Twitter network was 0.443 and the neighborhood assortativity was 0.689. Both the assortativity values were found to be statistically significant. The results showed that users group on Twitter according to their level of happiness.

While there are several works that shed light on the characteristics of the Twitter networks, the analysis of Twitter users from the perspective of languages has recently been explored by network scientists.

In order to understand the distribution of languages on Twitter, Hong et al. analyzed the languages of tweets posted by the users\footnote{6}. The authors demonstrated how the usage of Twitter characteristics such as URLs, hashtags, mentions, tweets, and retweets vary. In Table \ref{table:languages}, the authors showed the top ten languages according to the number of tweets. In their dataset, 95.6\% of the tweets were in the top ten languages. English was the most used language followed by Japanese, Portuguese, Indonesian and Spanish. The authors showed the distribution of the Twitter characteristics in Table \ref{table:twitter-characters}. Twenty one percent of the tweets contained URLs and 49\% contained mentions. Korean and Indonesian users retweeted as well as used mentions in
their tweets very frequently. German users used URLs and hashtags more than other language users. Weerkamp et al., who carried out a similar study, strengthened the findings of Hong et al. Weerkamp et al.’s work indicated that Indonesian users have a large number of mentions in their tweets and German tweets are very likely to contain hashtags as well as URLs. In fact, one in every four tweets in German contained hashtags.

Table 2.2: Top 10 languages on Twitter according to the number of tweets (source: Hong et al. [6]).

<table>
<thead>
<tr>
<th>Language</th>
<th>Tweets</th>
<th>%</th>
<th>Users</th>
<th>Tweets/user</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>31,952,964</td>
<td>51.1</td>
<td>5,282,657</td>
<td>6</td>
</tr>
<tr>
<td>Japanese</td>
<td>11,975,429</td>
<td>19.1</td>
<td>1,335,074</td>
<td>9</td>
</tr>
<tr>
<td>Portuguese</td>
<td>5,993,584</td>
<td>9.6</td>
<td>993,083</td>
<td>6</td>
</tr>
<tr>
<td>Indonesian</td>
<td>3,483,842</td>
<td>5.6</td>
<td>338,116</td>
<td>10</td>
</tr>
<tr>
<td>Spanish</td>
<td>2,931,025</td>
<td>4.7</td>
<td>706,522</td>
<td>4</td>
</tr>
<tr>
<td>Dutch</td>
<td>883,942</td>
<td>1.4</td>
<td>247,529</td>
<td>4</td>
</tr>
<tr>
<td>Korean</td>
<td>754,189</td>
<td>1.2</td>
<td>116,506</td>
<td>6</td>
</tr>
<tr>
<td>French</td>
<td>603,706</td>
<td>1.0</td>
<td>261,481</td>
<td>2</td>
</tr>
<tr>
<td>German</td>
<td>588,409</td>
<td>1.0</td>
<td>192,477</td>
<td>3</td>
</tr>
<tr>
<td>Malay</td>
<td>559,381</td>
<td>0.9</td>
<td>180,147</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 2.3: Percentage of characteristics on Twitter. (source: Hong et al. [6]).

<table>
<thead>
<tr>
<th>Language</th>
<th>URLs</th>
<th>Hashtags</th>
<th>Mentions</th>
<th>Replies</th>
<th>Retweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>21%</td>
<td>11%</td>
<td>49%</td>
<td>31%</td>
<td>13%</td>
</tr>
<tr>
<td>English</td>
<td>25%</td>
<td>14%</td>
<td>47%</td>
<td>29%</td>
<td>13%</td>
</tr>
<tr>
<td>Japanese</td>
<td>13%</td>
<td>5%</td>
<td>43%</td>
<td>33%</td>
<td>7%</td>
</tr>
<tr>
<td>Portuguese</td>
<td>13%</td>
<td>12%</td>
<td>50%</td>
<td>32%</td>
<td>12%</td>
</tr>
<tr>
<td>Indonesian</td>
<td>13%</td>
<td>5%</td>
<td>72%</td>
<td>20%</td>
<td>39%</td>
</tr>
<tr>
<td>Spanish</td>
<td>15%</td>
<td>11%</td>
<td>58%</td>
<td>39%</td>
<td>14%</td>
</tr>
<tr>
<td>Dutch</td>
<td>17%</td>
<td>13%</td>
<td>50%</td>
<td>35%</td>
<td>11%</td>
</tr>
<tr>
<td>Korean</td>
<td>17%</td>
<td>11%</td>
<td>73%</td>
<td>59%</td>
<td>11%</td>
</tr>
<tr>
<td>French</td>
<td>37%</td>
<td>12%</td>
<td>48%</td>
<td>36%</td>
<td>9%</td>
</tr>
<tr>
<td>German</td>
<td>39%</td>
<td>18%</td>
<td>36%</td>
<td>25%</td>
<td>8%</td>
</tr>
<tr>
<td>Malay</td>
<td>17%</td>
<td>5%</td>
<td>62%</td>
<td>23%</td>
<td>29%</td>
</tr>
</tbody>
</table>

Poblete et al. investigated the languages of the tweets of users with respect to their countries. The authors examined the top ten countries according to the number of users as well as activity rates on Twitter as shown in Figure 2.7. The authors also investigated the top three languages in the top ten countries, and represented in Figure 2.8. English was found to be
the most popular language. According to the authors, more than 10% of the tweets in English were from users who were from the Netherlands, Mexico, and Indonesia and 9% of the tweets in English were from users who were from Brazil. One interesting finding in this work was that Italian was among the commonly used languages in the top ten countries although Italy was not among the top ten countries according to the number of users.

Figure 2.7: Distribution of users of the top ten countries on Twitter (source: Poblete et al. [4]).

Figure 2.8: Distribution of top three languages in the top ten countries on Twitter (source: Poblete et al. [4]).
Nguyen et al. studied the relation between language and age of the Dutch users on Twitter. It was found that the use of capitalized words, word length, tweet length, links, and hashtags on Twitter can indicate the age of Dutch users \[36\]; the aforementioned tweet characteristics vary to a great extent among young users and are quite stable among older users. Older users use longer tweets, as well as more links and hashtags. Younger people use more alphabetical lengthening (for example, niiiiiiice instead of nice) and capitalized words (for example, HAHA, LOL).

Java et al. performed an analysis on a Twitter follower network (with reciprocal connections) to understand the influence of languages in the social network of users within the continents (Asia, Europe, Oceana, North America, South America, and Africa) \[37\]. Users from Europe and Asia tend to have high reciprocity, clustering coefficient, and degree correlation in their subgraphs. The authors suggested that users from Japan and the Spanish speaking countries generally connect to others who use the same language, which probably explains the high reciprocity and clustering coefficients of the networks.

Mocanu et al. demonstrated the linguistic diversity on scales ranging from country-level aggregation to city-level neighborhoods on Twitter \[5\]. The authors found that the tweets generated within a country were usually written in the dominant language of the country. Users from non-English speaking countries tweet in English as a way to reach out to more users on Twitter. The authors performed city-level analysis on Montreal and New York City (NYC). In Montreal, a spatial segregation was observed between the English and the French users. The French users were mostly concentrated in the northern neighborhoods of Montreal. Figure 2.9 shows the distribution of the English and the French users in Montreal. In NYC, the location of the non-English users was prominent. The authors found distinct Spanish communities around Harlem, the Bronx and Queens as well as Korean communities around Palisades Park, NJ, and Flushing, NY. Another key finding of Mocanu et al. was the reason behind the heterogeneity of the Twitter adoption in countries. The authors suggested that economic diversity of the countries is the primary source of the heterogeneity of Twitter adoption. The economic diversity of a country was captured using the Gross Domestic Product (GDP). The GDP per capita of countries was found to be positively correlated with the number of users per capita.

Ronen et al. measured the global influence of a language based on its position in the language network generated from the tweets of users, Wikipedia editors, and book translations \[38\]. In
the Wikipedia language network, every node represented a language and every edge between the two languages represented a Wikipedia page that was edited in the two languages. In the Twitter language network, every node represented a language and every edge between the two languages indicated that a user who tweeted in one language, also tweeted in another. In the book translation language network, nodes represented the languages and the edges indicated that a book was translated from one language to another. The three language networks, although representative of different activities of users, were similar in terms of the strength of the links. The authors argued that the position of a language in the language network can affect the spread of information in the network. Better connected languages increase the chances of faster diffusion of information. Another finding of the researchers was that the position of a language in a language network is related to the cultural influence of its users. The authors showed that the eigenvector centrality of the languages is positively correlated with the number of famous people associated with the language. The famous people in the datasets included elites such as Einstein, Darwin, and Picassos. The languages of the elites were the languages associated with their country of birth.
The BBC reported that there are approximately 7,000 spoken languages in the world, 90% of which are spoken by fewer than 100,000 people \[39\]. More than one million people communicate in 150-200 languages and 45 languages have only a single speaker. The cause of the drop in the number of spoken languages has been an area of interest to linguists and other related research communities. One of the popular proposals that has long been debated tries to intertwine languages with the global economies and indicated that economically strong nations culturally influence the nations that do not have sound economies \[40, 41\]. Scientists argue that English skills of a population and economic performance of a country are directly correlated \[42, 43\]. In an interesting study by Ranis, it is found that human development has important effects on economic growth \[44\]. An increase in the capabilities available to individuals allows them to pursue occupations in which they are more productive. In this sense, human development is correlated with the human capital and human capital in turn correlates with economic growth. Thus, human development has an effect on the economic growth. Kulshrestha et al. demonstrated that human development is highly correlated with Twitter adoption in countries \[45\]. Furthermore, the authors also demonstrated that the US accounts for 25% of the world GDP and 72% of all tweets produced in Twitter.

While there are several works that have explored the characteristics of the Twitter networks and the distribution of languages on Twitter, only a few works paid attention to the regularities that emerge from patterns of the languages used by these users and how the users organize themselves according to their language preferences \[5, 38\]. Additionally, the relation between the human development and the language preferences of users on Twitter also lacks investigation.

### 2.4 Literature Gap

Despite the vast literature on the language analysis in Twitter, the research community has not addressed a few important questions related to the language analysis on Twitter. In this section, we summarize a few questions that this dissertation investigates. In the following chapters, we discuss our work in detail.
2.4.1 Problem Statement

- **Question 1:**

  Can languages act as a catalyst in the formation of user connections on Twitter?

  User attributes such as gender, religion, and political views play a role in the formation of connections; people tend to link to others who share similar views. If the connectivity of users are restricted to their languages, novel information may not diffuse in the network. Keeping this in mind, it is interesting to investigate how users organize on Twitter according to their languages.

- **Question 2:**

  How different are the language connections in different Twitter activities?

  The language relations extracted from the retweet and the follower networks may vary. The preference of users to connect may be significantly different when one tries to receive information (as in following another user) with respect to when one tries to transmit information (as in retweeting). A user may be willing to receive information in another language but not likely to pass it on. Since Twitter is a free platform to communicate, the language associations are a lot more about the language relationships today than historical accounts.

- **Question 3:**

  Does the position of a language in a network play any role in users’ exposure to information?

  Diffusion of information in all the languages may not be equal. A few languages hold advantageous positions in the network. For example, we know that information may diffuse faster and wider in English than in Swahili. Users of the most connected languages can spread information in the network with less difficulty, whereas other language users may have difficulty in spreading it. Only a handful of languages are globally very popular. It is possible that people who use popular languages have more friends, are exposed to more information, and are more likely to spread information.
• Question 4:

Can the characteristics of the languages manifest development-related outcomes of the society?

The social development of a nation is a multi-dimensional concept and not related to economic growth alone. Since languages are infused in every facet of our social lives, it is interesting to inspect if the languages can indicate user development.

2.4.2 Thesis Statement

Twitter users tend to demonstrate a strong language preference while passing messages among themselves. The network structure of languages formed by the multilingual users in Twitter indicate that only a few languages occupy central positions, and the positions of the languages in the network correlates significantly with the Human Development Index of the users.

This dissertation investigates the effect of language in the follower and the retweet networks of users on Twitter. We find that the preference of language among the users is much higher during retweeting than during following. The language structure formed by the multilingual users on Twitter indicates that only a handful of languages are popular. This language structure also shows us that the centralities of the languages significantly correlate with the Human Development Index of the users.
Chapter 3

Datasets and Network Validation

3.1 Datasets

Twitter officially supports 48 languages on its website, but if users use their phones, Twitter
is able to get the language configuration of the phone even if the language is not one of
the 48 supported directly. Android phones support 70 languages\(^1\) and iOS phones support
more than 40 languages\(^2\). We used an API to collect data from Twitter. We collected two
datasets using different approaches to make sure we minimized possible biases arising from one
dataset use. Given that extracting the entire Twitter data is impossible due to limitations
imposed by Twitter, we used one of the popular network sampling strategies discussed by
Hanneman et al.: the ego-centric (with alter connections) approach \(^46\). In the aforementioned
network sampling, we needed to identify a few focal nodes or egos (people in the social network)
and collect their alters (or friends). In many cases, it was not possible to track down all the alters
of the egos, so we collected some of them randomly. We derived some statistics of our sample
networks and show that they exhibit characteristics that are expected as a result of many of the
real-world networks studied in the past.

\(^1\)http://resources.globalizationpartners.com/blog/android-lollipop-and-multiple-languages
\(^2\)https://developer.apple.com/internationalization/
3.2 Network Generation

3.2.1 Dataset 1: Follower Network

The Indo-European language family has the largest number of speakers (45% of the world population) as well as the widest dispersion around the world \[47\]. We chose to work with English, German, Russian, and Spanish as starting languages for our data collection procedure because they have a large number of speakers from the Indo-European family.

We selected the top 20 famous egos of each of the four languages (English, Spanish, Russian, and German) by using the Carousel feature of Google \[48\], which displays results based on reviews, pictures, and numerous other factors. Table 3.1 shows the Twitter profile names of the egos in our dataset. Given that 93.6% of Twitter users have fewer than 100 followers \[10\], we collected a random sample of up to 100 followers of every ego (1-step-neighborhood) and up to 100 followers of every follower of the ego (2-step-neighborhood). We chose to keep the number of followers consistent in both the neighborhoods. We built 80 ego networks with their 1-step-neighborhood and 2-step-neighborhood followers. In our ego networks, every node represents a Twitter user and every edge represents a follower relationship. We used the Twitter API to look up the profile language of every user. Though the egos have their Twitter profile language as one of the four languages we started with, their followers may have different profile languages. When we collected the languages of the followers, we were able to have a representation of most of the languages currently available on Twitter (57 languages). Figure 3.1 shows the sample network of one of the egos in our dataset: Vladimir Putin (with his 1-step and 2-step followers), the current President of the Russian Federation. We merged all the ego networks to form a single Twitter follower network. Our follower network consists of 170,082 nodes and 237,588 edges. It is to be noted that we did not generate our follower network based on the reciprocal relationships, where egos and alters follow each other, unlike many previous works \[31\], \[1\]. If such relations happened to be in our network, we also included it in our study. Since, in general, very few users have reciprocal relationships, we decided to keep our network as a small representation of the Twitter network. Figure 3.2 depicts a sample of the follower network of a few Russian egos. We colored the nodes based on their languages. The size of a node represents the number of followers of the user.
Next, we used the language information of every user in the follower network to generate our *follower language network (FLN)*. In the FLN, every node is a language as depicted in Figure 3.3. Two languages are connected if two connected users (in the follower network) use different languages. The color of the nodes represents the betweenness, and the size of the nodes represents their degree. In the FLN, we have 57 languages represented.
Table 3.1: List of egos per language in the Follower Network dataset.

<table>
<thead>
<tr>
<th>English</th>
<th>German</th>
<th>Russian</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>@bflay</td>
<td>@ElverfeldSven</td>
<td>@5umm</td>
<td>@albertochicote</td>
</tr>
<tr>
<td>@Bourdain</td>
<td>@grandcheflafer</td>
<td>@achekhova</td>
<td>@eneko_atxa</td>
</tr>
<tr>
<td>@Chef_Keller</td>
<td>@HolgerStromberg</td>
<td>@ARTEM_KLYUSHIN</td>
<td>@evaarguinano</td>
</tr>
<tr>
<td>@ChefChiarello</td>
<td>@SchultheckAlfons</td>
<td>@borisskumin</td>
<td>@JordiCruzMas</td>
</tr>
<tr>
<td>@tomcocolchio</td>
<td>@traubetonbach</td>
<td>@MedvedevRussia</td>
<td>@karguinano</td>
</tr>
<tr>
<td>@jtimberlake</td>
<td>@Helene_Fischer</td>
<td>@mnzadornov</td>
<td>@AlejandroSanz</td>
</tr>
<tr>
<td>@kanyewest</td>
<td>@JeanBiedermann</td>
<td>@navalny</td>
<td>@antoniofanderas</td>
</tr>
<tr>
<td>@katyperry</td>
<td>@naidooxavier</td>
<td>@Nyusha_Nyusha</td>
<td>@BardemOficial</td>
</tr>
<tr>
<td>@ladygaga</td>
<td>@nenaofficial</td>
<td>@PutinRF</td>
<td>@ma silicastr</td>
</tr>
<tr>
<td>@taylorswift13</td>
<td>@sarahconnorfc</td>
<td>@RealVolya</td>
<td>@mario casas</td>
</tr>
<tr>
<td>@arwa_journalist</td>
<td>@BettinaBelitz</td>
<td>@urgantcom</td>
<td>@AbeInfanzon</td>
</tr>
<tr>
<td>@EmbJournalist</td>
<td>@GuentherJauch65</td>
<td>@VictoriaBorya</td>
<td>@ConchaGarciaCam</td>
</tr>
<tr>
<td>@Jeff_Journalist</td>
<td>@Rabipeter</td>
<td>@ViktorPelevin</td>
<td>@LettThePrincess</td>
</tr>
<tr>
<td>@JournalistFritz</td>
<td>@Sky_Lierhaus</td>
<td>@vladimirpozner</td>
<td>@lopezgarrido</td>
</tr>
<tr>
<td>@JournalistJohnM</td>
<td>@UdoUlfkotte</td>
<td>@Zhirinovskiy</td>
<td>@pedrojramirez</td>
</tr>
<tr>
<td>@BillSimmons</td>
<td>@Fanorakel</td>
<td>@bilan</td>
<td>@albertocontador</td>
</tr>
<tr>
<td>@DavyaneWade</td>
<td>@FelixLoch</td>
<td>@fkrkorov</td>
<td>@alo_oficial</td>
</tr>
<tr>
<td>@KDTray5</td>
<td>@Katarina_Witt</td>
<td>@JuliaVolkova</td>
<td>@CasillasWorld</td>
</tr>
<tr>
<td>@King.James</td>
<td>@ThomasHitz</td>
<td>@Masha</td>
<td>@lorenzo09</td>
</tr>
<tr>
<td>@SHAQ</td>
<td>@timoboll</td>
<td>@Orbakaite</td>
<td>@RafaelNadal</td>
</tr>
</tbody>
</table>

3.2.2 Dataset 2: Retweet Network

Our second dataset was a collection of tweets about the leaders of the Group of Twenty\(^3\) for a period of 35 days. The Group of Twenty (G-20), which was founded in 1999, is an international organization of economic cooperation and decision making. The G-20 involves 19 individual countries plus the European Union and accounts for 85% of the world gross product, 80% of the world trade, as well as two-thirds of the world population. We collected 10,610,653 tweets that consisted of one or more of the last names (in English) of the leaders of the G-20\(^4\). In Table 3.2, we represent the last names of the leaders in our retweet dataset. Figure 3.4 shows a snapshot of tweets that contain the last names of leaders from the G20. Although there could be a few limitations to our data collection method, we are more interested in deriving insights about the language relations on Twitter; hence name inconveniences do not affect the study. By capturing the tweets about the G20 leaders, we were able to capture a great deal of language diversity in Twitter. The users who posted the tweets were expected to use different languages.

\(^3\)http://g20.org

26
We collected the user of every retweet and the user of the original tweet. We used the information to build a network called *retweet network*. We also collected the language of these users on Twitter. Our retweet network consists of 1,922,815 nodes and 4,131,866 edges. In this retweet network, every node represents a user and every edge between the two users represents that they retweeted each other.

We extracted the language network from the retweet network and named it *retweet language network (RLN)*. Every node in our RLN is a language. Two languages in the RLN are connected if two connected users in the retweet network use different profile languages. We have 51 languages.
Figure 3.3: FLN. The color of a node represents the betweenness of the node in the network: color ranges from green (low betweenness) to red (high betweenness). The size of a node represents degree of the node. The width of an edge represents the weight of the edge. In the network above, most of the information passes through English (en), followed by Spanish (es), Italian (it), Russian (ru), German (de), Portuguese (pt), French (fr), Swedish (sv), Turkish (tr), and Indonesian (id). The aforementioned languages are also very strongly connected in the network.

In Table 3.3, we summarize the statistics of our follower and retweet networks. It is to be noted that we demonstrate the characteristics of the sample Twitter data with a limited number of celebrities and followers as well as a limited number of tweets. As mentioned before, although it is very interesting to analyze the characteristics of the whole Twitter network with all the users, it is something Twitter does not provide with its API. If the entire Twitter network is considered, it is likely that the average path length of the networks would be smaller because the networks would be more dense. After we constructed our Twitter networks, we were also curious to find out the characteristics of other existing social networks. We listed a few social
Table 3.2: List of last names of the G20 leaders in the Retweet Network dataset.

<table>
<thead>
<tr>
<th>Track words</th>
<th>Track words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zuma</td>
<td>Obama</td>
</tr>
<tr>
<td>Kirchner</td>
<td>Jinping</td>
</tr>
<tr>
<td>Roussef</td>
<td>Noda</td>
</tr>
<tr>
<td>Pena Nieto</td>
<td>Myung-bak</td>
</tr>
<tr>
<td>Harper</td>
<td>Singh</td>
</tr>
<tr>
<td>Cameron</td>
<td>Gillard</td>
</tr>
<tr>
<td>Erdogan</td>
<td>Abdullah</td>
</tr>
<tr>
<td>Merkel</td>
<td>Hollande</td>
</tr>
<tr>
<td>Renzi</td>
<td>Putin</td>
</tr>
<tr>
<td>Yudhoyono</td>
<td></td>
</tr>
</tbody>
</table>

CatarinaTyylee: I said I miss Chris but he will come back and I can’t wait for that day & Cameron specially is going to be so happy 😄

TrumanTown: said RT @libertea2012: How Obama Could Help 6.1 Million Workers with a Stroke of His Pen: The Fair Labor Standards... http://t.co/6VpxAy9yQh #p2...

jordynspirls: Said What if Cameron actually can sing, and one day he randomly comes out with an album.

tadduncan: said RT @smarty216b: Michael Sam scores Johnny Manziel. #ESPN 10-part film series to debut Monday chronicling history of the play. President Obama.

arctony: said RT @tomInPSM: #Obama orders review of police use of military equipt http://t.co/xr06VIEEt7 via @usat hiptopnews #ferrygates #WhiteHouse http://–

kingpayton: said RT @sarah_danielle1: Obama rejected the ALS challenge... George W. Bush wore a sock and dumped dice on his head... #america

Figure 3.4: Snapshot of the tweets that contain the last names of leaders of the G20, that are posted by Twitter users

networks in Table 3.4 that are formed by people in different areas and are often referred to by the researchers in network science [50]. These social networks represent canonical datasets that are frequently studied by researchers to understand the characteristics of networks. The sizes of the networks vary, ranging from only 2,018 nodes in the Protein interaction network to 702,388 nodes in the Actor network. In addition, a few networks are directed while other are undirected. The sizes and the characteristics of the reference social networks in Table 3.4 also give us confidence that although our Twitter networks are small, they could be successfully used to understand the characteristics.
Table 3.3: Description of the two datasets used in this work. A third dataset is used in Chapter 6 only and its description is included in that chapter.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Dataset 1 (Follower Network)</th>
<th>Dataset 2 (Retweet Network)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start Data Collection</td>
<td>January, 15, 2015</td>
<td>August, 24, 2014</td>
</tr>
<tr>
<td>Finish Data Collection</td>
<td>February, 15, 2015</td>
<td>September, 29, 2014</td>
</tr>
<tr>
<td>Number of Days</td>
<td>32</td>
<td>35</td>
</tr>
<tr>
<td>Number of Nodes</td>
<td>170,082</td>
<td>1,922,815</td>
</tr>
<tr>
<td>Number of Edges</td>
<td>237,588</td>
<td>4,131,866</td>
</tr>
<tr>
<td>Transitivity</td>
<td>0.018</td>
<td>0.003</td>
</tr>
<tr>
<td>Mean Degree</td>
<td>2.864</td>
<td>4.299</td>
</tr>
<tr>
<td>Average Path Length</td>
<td>3.356</td>
<td>9.953</td>
</tr>
<tr>
<td>Language network extracted</td>
<td>FLN</td>
<td>RLN</td>
</tr>
<tr>
<td>Languages in the language network</td>
<td>57</td>
<td>51</td>
</tr>
</tbody>
</table>

Table 3.4: Description of the networks that exist in our society and have been studied previously. They differ widely in their sizes. \( N \) refers to the number of nodes, \( L \) refers to the number of edges and \( <k> \) refers to the mean degree of the networks.

<table>
<thead>
<tr>
<th>Network</th>
<th>Nodes</th>
<th>Links</th>
<th>Directed/Undirected</th>
<th>( N )</th>
<th>( L )</th>
<th>( &lt;k&gt; )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobile Phone Calls</td>
<td>Subscribers</td>
<td>Calls</td>
<td>Directed</td>
<td>36,595</td>
<td>91,826</td>
<td>2.51</td>
</tr>
<tr>
<td>Email</td>
<td>Email addresses</td>
<td>Emails</td>
<td>Directed</td>
<td>57,194</td>
<td>103,731</td>
<td>1.81</td>
</tr>
<tr>
<td>Science Collaboration</td>
<td>Scientists</td>
<td>Co-authorships</td>
<td>Undirected</td>
<td>23,133</td>
<td>93,437</td>
<td>8.08</td>
</tr>
<tr>
<td>Actor Network</td>
<td>Actors</td>
<td>Co-acting</td>
<td>Undirected</td>
<td>702,388</td>
<td>29,397,908</td>
<td>83.71</td>
</tr>
<tr>
<td>Citation Network</td>
<td>Papers</td>
<td>Citations</td>
<td>Directed</td>
<td>449,673</td>
<td>4,689,479</td>
<td>10.43</td>
</tr>
<tr>
<td>Protein Interactions</td>
<td>Proteins</td>
<td>Binding interactions</td>
<td>Undirected</td>
<td>2,018</td>
<td>2,930</td>
<td>2.90</td>
</tr>
</tbody>
</table>
Figure 3.5: RLN. The color of a node represents its betweenness in the network: color ranges from green (low betweenness) to red (high betweenness). The size of a node represents the degree of the node. The width of an edge represents the weight of the edge. In the network above, most of the information passes through English (en), followed by Spanish (es), Italian (it), Russian (ru), German (de), Portuguese (pt), French (fr), Swedish (sv), Turkish (tr), Arabic (ar), Chinese (zh) and Dutch (nl). The aforementioned languages are also very strongly connected in the network.
3.3 Network Data Validation

In statistical analysis, Hypothesis Test is done to test the validity of a claim. In the Hypothesis Test, null hypothesis refers to a claim regarding a population and alternative hypothesis refers to be the one that is to be believed, if the null hypothesis is untrue. Statistical p-value determines the significance of the test. A small p-value (usually less than 0.05) indicates strong evidence to reject the null hypothesis and accept the alternative hypothesis. A large p-value (usually greater than 0.05) indicates weak evidence against the null evidence, and therefore fail to reject the null hypothesis. For all of our statistical validations, we used a p-value < 0.05 to determine the significance of our results.

Since the datasets were collected using two different approaches, we examined if our language networks (FLN and RLN) are similar. Although the language relations in each of these networks represent different user activities on Twitter, they have similar characteristics in particular to the context of weights between two languages. We computed the correlation of every pair of languages in both the networks. We found a positive correlation of $r=0.91$ (with p-value < 0.05). The correlation indicates that the language relations are somewhat invariant across both the networks; this gives us more confidence that the relations extracted are meaningful because they come from two different forms of interactions between users.

Table 3.5: The log-likelihood comparative study shows that log-normal best describes the in-degree and out-degree distributions of the follower network.

<table>
<thead>
<tr>
<th>Candidate Distributions</th>
<th>In-degree Fitting (likelihood, p)</th>
<th>Candidate Distributions</th>
<th>Out-degree Fitting (likelihood, p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Power-law, Log-normal)</td>
<td>(-358.22, p &lt; 0.05)</td>
<td>(Power law, Log-normal)</td>
<td>(-2,482.09, p &lt; 0.05)</td>
</tr>
<tr>
<td>(Power-law, Truncated Power law)</td>
<td>(-152.59, p &lt; 0.05)</td>
<td>(Power law, Truncated Power law)</td>
<td>(-2,220.68, p &lt; 0.05)</td>
</tr>
<tr>
<td>(Power-law, Exponential)</td>
<td>(17,046.39, p &lt; 0.05)</td>
<td>(Power law, Exponential)</td>
<td>(-2,324.18, p &lt; 0.05)</td>
</tr>
<tr>
<td>(Log-normal, Exponential)</td>
<td>(17,404.61, p &lt; 0.05)</td>
<td>(Log-normal, Exponential)</td>
<td>(157.90, p &lt; 0.05)</td>
</tr>
<tr>
<td>(Log-normal, Truncated Power law)</td>
<td>(205.63, p &lt; 0.05)</td>
<td>(Log-normal, Truncated Power law)</td>
<td>(261.41, p &lt; 0.05)</td>
</tr>
</tbody>
</table>

It is important to first analyze the similarity of weights of the language relations in both the language networks because our interest lies in the investigation of the characteristics of the languages. Once we found that the language networks that we extracted have meaningful relations, we investigated the follower and the retweet networks as well.
We generated in-degree as well as out-degree distributions of the follower and the retweet networks and assessed the goodness-of-fit by comparing with other distributions using log-likelihood [51]. The comparative study helped us to identify the best possible fitting that described the degree distribution. We compared the degree distributions with log-normal, power-law, truncated power law, and exponential because these are the common distributions found in real-world networks. Table 3.5 and Table 3.6 show the analysis of the degree distributions of our follower network and retweet network respectively. The “Candidate Distributions” column shows the distributions that were tested to find out the best fit of our data. The corresponding value of the fit is represented by a log-likelihood ratio column with a significance value (p-value). The log-likelihood value is positive if the data is more likely to fit the first distribution and negative if the data is more likely to fit the second distribution. The statistical significance of the fit is given by the p-value. When the in-degree and the out-degree distributions were compared against all the possible combinations of the four distributions we tested, we found that the power-law is not the best fit (as discussed by Kwak et. al [31]). We also represent the degree distribution fits of the follower and retweet networks in Figure 3.6 and Figure 3.7. In addition, we show that the log-normal is a better fit of the degree distributions with a significance level of
p < 0.05. Significant research has been done on online social networks that exhibit log-normal distributions [52,53].

Table 3.6: The log-likelihood comparative study shows that log-normal best describes the in-degree and out-degree distributions of the retweet network.

<table>
<thead>
<tr>
<th>Candidate Distributions</th>
<th>In-degree Fitting</th>
<th>Out-degree Fitting</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Power law, Log-normal)</td>
<td>(-739.49, p &lt; 0.05)</td>
<td>(Power law, Log-normal)</td>
</tr>
<tr>
<td>(Power law, Truncated Power law)</td>
<td>(-693.89, p &lt; 0.05)</td>
<td>(Power law, Truncated Power law)</td>
</tr>
<tr>
<td>(Power law, Exponential)</td>
<td>(388,404.99, p &lt; 0.05)</td>
<td>(Power law, Exponential)</td>
</tr>
<tr>
<td>(Log-normal, Exponential)</td>
<td>(389,144.48, p &lt; 0.05)</td>
<td>(Log-normal, Exponential)</td>
</tr>
<tr>
<td>(Log-normal, Truncated Power law)</td>
<td>(45.59, p &gt; 0.05)</td>
<td>(Log-normal, Truncated Power law)</td>
</tr>
</tbody>
</table>
Figure 3.7: Comparison of the distributions demonstrating that the log-normal distribution best describes the degree distribution of the retweet network. Table 3.6 shows the comparison results.
Chapter 4

Language-Centric User Connectivity Analysis

Given that languages are embedded in our everyday lives, it is particularly interesting to explore how they (languages) influence and shape us in our society. The native language (mother tongue) of an individual is of prime importance to group formation; language together with religion and skin color has been shown to be of great significance to the patterns of interactions within societies [54]. Note that in Section 2.2 of Chapter 2, we showed that assortativity captures the tendency of users to connect to similar others (in terms of gender, religion, and political views) [23]. There has been evidence that such propinquity to connect to alike may be beneficial or detrimental to those who are depending on the particular situation [23]. It is therefore not surprising if we find that users tend to connect to others who use same language on Twitter; language is expected to be a catalyst for the formation of connections. A common language may lead to more efficient interactions. Yet, to our knowledge, the extent of the role played by language in connection formation on Twitter has not received much attention from the research community. In order to achieve an understanding of the importance of languages on user interactions, we studied the structure of connectivity of several users as a function of their languages on Twitter. Recall some of the questions that we addressed in Chapter 1 e.g., “Are there different patterns
of connectivity for users of different languages?”, and “Are users organized according to their languages?” In this chapter, we present our findings related to the aforementioned questions.

Our first analysis was to investigate the popularity of languages on Twitter. We collected the total number of users in every language in our datasets; Figure 4.1 shows the twenty most popular languages in our follower and retweet networks. We found that the top ten languages are the same in both networks (with the exception of Arabic and Indonesian). However, the order of popularity of the languages varies. Arabic ranks eighth in the follower network and eleventh in the retweet network. On the other hand, Indonesian ranks twelfth in the follower network and ninth in the retweet network. In both networks, English has the highest number of users, followed by Spanish. Other languages such as German, Russian, French, Portuguese, Italian, Arabic, Turkish, Japanese, and Indonesian are also prominently used on Twitter and occupy intermediary positions in the rankings. It is important to note that our ranking of languages does not reflect the ranking of the real-world or the rankings of Twitter languages as reported by Statista [55]. In real-world, Ethnologue reports that Chinese is the most spoken language in the world [56]. Our analysis is only based on Twitter. Twitter is not very popular in Russia and blocked in China. Twitter penetration in the countries is one of the major reasons for some of the disparities. However, the disparities do not hinder extracting insights about the users in countries where Twitter is widely used. Our rankings are also different than what we have found in literature [31]. One possible explanation could be that our analysis is very recent in comparison to the other studies. Further analysis could be done to understand the dynamics of these language distributions (how the distribution is changing); however, we have it as our future work.
4.1 Assortativity of Users

4.1.1 Degree Assortativity

As mentioned before, degree assortativity is the tendency of nodes to connect to others with a similar degree; nodes that have a high degree tend to connect to others with a high degree and the nodes that have a low degree tend to connect to others with a low degree. We measured the degree assortativity of the networks using the equation described in Subsection 2.2.8 of Chapter 2. The degree assortativity coefficient of our follower network is -0.18, which indicates that the network is disassortative by degree. Figure 4.2 shows that the nodes that have a low degree are connected to nodes that have a high degree in the follower network. Furthermore, the negative assortativity
of the follower network means that the users in Twitter do not connect to others with similar degrees. They usually tend to connect to those who are already popular in Twitter (have a high degree). For example, users usually follow their favorite celebrities. Following enables the users to receive information posted by the celebrities. On the other hand, celebrities do not connect to other celebrities, probably due to competition between them for followers. The same is true for low degree users who do not connect to other low degree users. Twitter may very well represent a broadcasting platform where users like to connect to people who are sources of “exciting” information.

Figure 4.2: Illustration that the nodes with a low degree have a high average neighbor degree and vice versa in the follower network, indicating disassortativity

The degree assortativity coefficient of the retweet network is $-0.06$. In Figure 4.3 we show that the nodes with a low degree have neighbors with a high degree and the nodes with a high degree have neighbors with a low degree. Such observation is probably because there are few users who create and post interesting tweets. The interesting tweets get retweeted many times and become popular. The users who create interesting tweets become sources of information. Most of the users usually retweet the information to pass it on to followers.
4.1.2 Language Assortativity

To estimate the extent to which a user tends to connect to another user who uses his language, we used the pairwise connection of users in relation to the languages they use on Twitter. We measured the assortativity coefficient of our follower network and the retweet network according to the language of the users; note the assumption here is that the user language is the preferred language of interaction of the user (previously mentioned). The language assortativity coefficient of the follower network is 0.56, which means that Twitter users display a strong preference for following people who use the same language as them. The assortativity coefficient of the retweet network is 0.74, which means that Twitter users also display strong association by languages while retweeting. However, the difference between one and the other is quite significant and indicates that despite the user’s preference to connect to others with the same language, the preference is stronger when we look at the information that is being transmitted (retweet). This is a strong indication that language plays a crucial role in information spread. Users are a lot more likely to pass on a tweet in the language they prefer than just follow another user in that same language. In other words, a user may be willing to receive information in a language different from his or
hers but not as likely to pass this information on (retweet the message). Table 4.1 summarizes the assortativity coefficients of both the networks.

Table 4.1: The table below shows that the networks are disassortative by degree and assortative by language.

<table>
<thead>
<tr>
<th>Assortativity</th>
<th>Follower network</th>
<th>Retweet network</th>
</tr>
</thead>
<tbody>
<tr>
<td>By Degree</td>
<td>-0.18</td>
<td>-0.06</td>
</tr>
<tr>
<td>By Language</td>
<td>0.56</td>
<td>0.74</td>
</tr>
</tbody>
</table>

The findings above demonstrate that Twitter users have a preference to connect to other users who use a common language, but more importantly, the preference is stronger while passing on a message than the trend of connecting to users with similar popularity. Hence, Twitter users are highly assortative by their languages but disassortative by their popularity (in terms of degree). However, we would like to point out that (i) user language may not necessarily be used as the de facto language in the tweet (further experiments need to be performed to understand if this is the case); (ii) celebrities and G-20 leaders in our case are popular in different areas, thus people with different interests put extra effort to understand the information they are “broadcasting” to the world; and (iii) the preference for who we follow is not transferred to the preference for who we use to get the information (retweets).

4.2 Diversity of Users

Assortativity of users does not capture the entire information in the sense that it gives the overall characteristic of the network and not the characteristic of the users of particular languages. Hence, it is necessary to understand the characteristics of users of every language in isolation.

We created a vector of language exposure of every user in the network. For example, if a network is formed by an English user A and his three friends who use Spanish, two friends who use German, three friends who use English (say, B, C and D) and one friend who uses Swahili (as depicted in Figure 4.5), then A’s language exposure can be represented by a vector $V^A_{English}$ such that the vector represents the number of friends of A according to languages. For better understanding, let us represent A’s language vector as $V^A_{English} = [3, 3, 1, 2]$, where, 

User Language Vector = [English, Spanish, Swahili, German]. Here, user A is exposed to four
languages. We generated the language exposure vector for every user in the network. Then we combined all the users of a particular language to have the language exposure vector of the particular language in the network. Considering the example above, let us say, we generated $V^A_{English}$, $V^B_{English}$, $V^C_{English}$ and $V^D_{English}$ to combine and form the English vector of the network, $V_{English}$. Likewise, we generated the exposure vectors for all the languages in the network. We normalized the vectors to have a unit vector of every language in the network. Recall that our follower network has users who use 57 different languages and our retweet network has users who use 51 different languages. One may argue here that while combining users to generate the language exposure vectors, higher degree users may influence the lower degree users. Note that it is unlikely to be the case in our study, because we limited our number of
followers. Next, using the language exposure vectors, we measured the extent of diversity of the languages in both of our networks. We measured the diversity of a language by calculating the ratio between the connections of the language to itself and the sum of its connections to other languages, as extracted from the language exposure vector. Figure 4.4(a) shows the diversity of languages in the follower network. We observe that some languages are more uniform (have fewer edges to other languages), or in other words, are less diverse. Languages like English (en), Spanish (es) or Russian (ru) usually exhibit a strong preference to connect within themselves in contrast to Georgian (ka), Urdu (ur) or Serbian (sr), which connect to other languages primarily. In the retweet network, we found that language preference is even more accentuated. Figure 4.4(b) shows that Turkish (tr), Italian (it), and English (en) mostly connect among themselves in comparison to languages like Lithuanian (lt), Serbian (sr), or Estonian (et).

Followed by the diversity analysis, we examined the patterns of connectivity of the languages by computing the entropy of the language exposure vectors. Note that this is important because
we could have a language that evenly connects to several other languages or we could have a language that heavily connects to only one language. Although in both the cases the languages are considered diverse, the diversity of the former language in the example is higher than the latter, or the association of the former language to other languages is much more disordered than the latter. In Figure 4.6(a), we demonstrate the entropy of languages in the follower network. Languages such as Greek (el) and French (fr) show high entropy, indicating that their association with other languages is very disordered. According to the French language exposure vector, French associates with itself as well as English almost evenly; Greek exhibits connections with several languages such as English, Spanish, Portuguese, German, and others. On the other hand, a few languages like Lithuanian (lt) and Malay (msa) have low entropy, meaning their association patterns are less disordered. Neither Lithuanian nor Malay exhibit any preference towards themselves, but Lithuanian shows a strong association with English and Malay shows a strong association with French. However, the retweeting and the follow patterns of the languages
vary. In the retweet network, as demonstrated in Figure 4.6(b), we observe that Malay (msa), Russian (ru), and Galician (gl) have very high entropy, which means their connecting patterns are disordered. Malay connects to several other languages such as Chinese, Japanese, English, Spanish, and French in a random fashion. On the other hand, languages with low entropy like Hindi (hi), Turkish (tr), and Italian (it) are less disordered in the network. Low entropy of a language may have two explanations. First, languages like Turkish and Italian show a very high preference to retweet among themselves. Second, languages like Hindi associate strongly with English although Hindi does not show a preference to connect with itself.

The association pattern of the languages on Twitter varies. Some languages show stronger association towards themselves, which may be because of the global popularity of the languages; users of such languages do not need to understand another language for information. Other language users show less association among themselves; such language users need to understand another language in order to receive information in the network. Due to the variation in the connection patterns, some languages are less disordered and some are more disordered on Twitter.
Chapter 5

Exploring the World Languages

Given that Twitter users prefer to connect to each other according to their languages, it is reasonable to assume that this preference may influence the flow of novel information. Therefore, to explore how information diffuses in Twitter, this chapter analyzes the highly-connected language network that underlies the social connectivity of the users. This language network is formed by the multilingual users who enable the spread of messages by becoming links between languages (recall the follower language network or FLN, and the retweet language network or RLN in Chapter 3). Unlike the social structure of friends, the structure of languages of friends is harder to be perceived because it emerges as a second level relation from users; instead of a connection between users, we connect the languages they use. Furthermore, not much effort has been made to understand the diffusion of information on Twitter and the role of languages in it [38]. In this chapter, we intend to (i) investigate the languages that are important in the information diffusion process, and (ii) determine the similarity of languages based on population preference. We have two main contributions in this chapter: (i) reachability of languages, which describes the connectivity of languages and shows how their positions in the network could affect their ability to send or receive information and (ii) similarity of languages, which demonstrates how languages are similar from the perspective of presence or absence of common neighbors (other languages) and also from the perspective of interactions of common users between the languages. We inspected our unweighted FLN and RLN networks to understand the reachability of languages and both the weighted as
well as the unweighted FLN and RLN for the similarity of languages. Just a reminder that in the unweighted language networks, we only considered the presence of the links and in the weighted language networks, we considered the weight of the links which represents the number of user interactions between the languages.

5.1 Reachability of Languages

Our first analysis was to estimate the density of the networks. The density of the FLN is 0.19, meaning 19% of all possible connections are present in the network and the density of the RLN is 0.27, meaning 27% of all possible connections are present in the network. The density measurements indicate that not all languages are well-connected in the networks. The connectivity of the languages can provide us a better understanding of their (languages) exposure to information. Languages that have a high degree are exposed to diverse information; therefore, they are more important in terms of receiving and spreading information. In our work, we consider the ten most connected languages as the important languages. We inspected the degree of all the languages to identify the important ones. Furthermore, the degree of the languages can show us the extent to which they can diffuse information on Twitter.

Table 5.1: Top ten languages by degree in the FLN and the RLN. The FLN has total 57 languages while the RLN has 51 languages represented.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Language</th>
<th>Degree</th>
<th>Language</th>
<th>Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>English (en)</td>
<td>53</td>
<td>English (en)</td>
<td>49</td>
</tr>
<tr>
<td>2</td>
<td>Spanish (es)</td>
<td>39</td>
<td>French (fr)</td>
<td>32</td>
</tr>
<tr>
<td>3</td>
<td>Russian (ru)</td>
<td>33</td>
<td>Spanish (es)</td>
<td>31</td>
</tr>
<tr>
<td>4</td>
<td>German (de)</td>
<td>32</td>
<td>Turkish (tr)</td>
<td>30</td>
</tr>
<tr>
<td>5</td>
<td>French (fr)</td>
<td>25</td>
<td>Russian (ru)</td>
<td>30</td>
</tr>
<tr>
<td>6</td>
<td>Italian (it)</td>
<td>25</td>
<td>Portuguese (pt)</td>
<td>29</td>
</tr>
<tr>
<td>7</td>
<td>Turkish (tr)</td>
<td>24</td>
<td>German (de)</td>
<td>26</td>
</tr>
<tr>
<td>8</td>
<td>Portuguese (pt)</td>
<td>24</td>
<td>Italian (it)</td>
<td>25</td>
</tr>
<tr>
<td>9</td>
<td>Indonesian (id)</td>
<td>22</td>
<td>Chinese (zh)</td>
<td>24</td>
</tr>
<tr>
<td>10</td>
<td>Japanese (ja)</td>
<td>21</td>
<td>Arabic (ar)</td>
<td>24</td>
</tr>
</tbody>
</table>

In Table 5.1, we show the top ten important languages in the FLN and the RLN. We found that in both the networks, English is the most connected language. A fad or gossip in English can spread much faster than in other languages through the multilingual users. The popularity
of English is well-known to everyone in today’s world. Our networks also reflect the same observation. Finding expected results intuitively gives us more confidence that the conclusions drawn from the network analysis are correct. Moreover, the FLN shows that French (fr) and Italian (it) as well as Turkish (tr) and Portuguese (pt) having the same degree. The RLN shows Turkish (tr) and Russian (ru) as well as Chinese (zh) and Arabic (ar) having the same degree. Note that one may ask whether two languages that have same degree play similar roles in the networks. To answer this question, we studied the structural similarity of the languages (see Section 5.2). The top languages in both the RLN and the FLN are almost the same, although the order of the languages varies. Why and how a language became important is part of our future work, although at this point we assume that it may be related to a number of factors such as immigration rates of the countries where they are used (that is, countries that have high immigration rates may have many users of several languages), historical background of the countries, academic standards, and more.

![Figure 5.1: Shortest paths between languages in the FLN. The diameter of the FLN is 4.](image)

The network of languages is quite complex. Languages that are directly connected may be able to exert more effect on each other than the ones that are not directly connected. Hence, we calculated the shortest paths between pairs of languages. Figure 5.1 shows the matrix
representation of the shortest paths in the FLN. The shortest path of a language to itself is zero. In addition, English is reached by almost all the languages in one hop except for az (Azerbaijani), gsw (Swiss German) and sw (Swahili). In fact, the three aforementioned languages are not directly connected to most of the languages in the network. Since the network is fully connected, a fad started in one language will eventually reach every other language, but it is likely that the transmission to the less connected languages will be slower. The FLN has diameter 4; no two languages in the FLN are more than four hops away from each other. Figure 5.2 shows the shortest paths in the RLN. We observe that the RLN has a language Lolcatz (lolc), which is an Internet meme language. Note that English connects to all the languages except Lolcatz (lolc) and Icelandic. The RLN has diameter 3.
5.2 Similarity of Languages

5.2.1 Similarity Based on the Existence of Connections

In our next analysis, we investigate if the languages can be grouped according to their connecting patterns. Grouping the languages can capture many characteristics of the languages, such as the possibility of future connections between languages and structural similarities of languages. There are several popular algorithms in Network Science that group entities. We used the nested stochastic block model algorithm (stochastic block model algorithm groups nodes in blocks with probabilities of connections between them [57] [58]) to find language groups. Figure 5.3 shows the application of the stochastic block model of our FLN. The languages are grouped into four blocks and are represented by different colors. The important languages are organized in the first and the second block. The languages that are connected to most of the important languages occupy the third block. The fourth block is formed by the least connected languages of the network. The tree representation in the FLN shows the subtrees at the first cutting level. The subtrees show the languages that belong to each block. The application of the stochastic block model of our RLN is represented in Figure 5.4. In the RLN, there are three blocks. The first block is occupied by the important languages. The second block is formed by the languages that connect to the important languages and receive intermediary attention in the network. The third block is formed by the least connected languages. The tree structure in the RLN shows the subtrees at the first cutting level and the subtrees show the languages that belong to each of the blocks.

Although the blocks indicate vital information about the languages, such as the languages that are similar according to their connecting patterns, what is more interesting is to quantify the similarity of the languages in Twitter. To measure the similarity between languages, we computed the Jaccard similarity index of every pair of languages. The Jaccard index is a metric that is used to capture similarity between two sets. It is defined as the ratio of the number of attributes shared between two sets divided by the total number of attributes present in either of them [59]. Jaccard’s index can be computed as follows:
Figure 5.3: FLN: the size of a node is based on its degree and the color of the node represents the block it belongs to. The nodes in the same block display similar pattern of connections. The tree representation in the FLN shows that there are four subtrees at the first cutting level. The subtrees show the languages that belong to each block.

\[ J = \frac{C}{A + B - C}, \]  

(5.1)

where \( C \) is the number of common attributes, \( A \) and \( B \) are the number of attributes present in each set. We depicted the similarity matrix of the FLN in Figure 5.5. Hindi (hi), Galician (gl), Lithuanian (lt), and Urdu (ur) have very high Jaccard coefficients. They are structurally similar because all of them connect to the important languages such as English, Spanish, and German in one step. So if Galician has a piece of information, Lithuanian is likely to have that information too. We also found that Estonian (et), Bengali (bn), Khmer (km), Irish (ga), Icelandic (is), Georgian (ka), and Bosnian (bs) have high Jaccard similarity coefficients because they are only connected to English in the network. Moreover, the important languages show high similarity. Figure 5.6 shows the similarity matrix of the RLN. Lithuanian (lt), Estonian
Figure 5.4: RLN: the size of a node is based on its degree and the color of the node represents the block it belongs to. The nodes in the same block have similar patterns of connections. The tree structure in the RLN shows that there are three subtrees at the first cutting level; that is, the important languages including English form a block and the other languages are arranged in two other blocks. The subtrees show the languages of every block.

(\textit{et}), Bengali (\textit{bn}), Marathi (\textit{mr}), Mongolian (\textit{mn}), Macedonian (\textit{mk}), Latvian (\textit{lv}), and Bosnian (\textit{bs}) display structural similarity due to their connection with English only. Like the FLN, the important languages in the RLN also show high structural similarity among themselves. Additionally, a few other languages also display structural similarity, such as Urdu–Slovak and Lloctaz–Icelandic.

### 5.2.2 Similarity Based on the Weight of Connections

Though our work examined many aspects of the language connections; the weight of the connections has not been captured so far. In our next analysis, we investigate how the languages can influence each other from the perspective of their weight of connections \( 60 \). Recall that
languages that have a higher weight indicate that there are more common users. The higher the number of common users between two languages, the higher is the likelihood that the information will flow from one language to another. Two strongly connected languages (having a higher edge weight) can diffuse information among each other faster than two weakly connected languages (having a lower edge weight).

To explore the similarity of languages based on their strength, we generated and normalized the weighted adjacency matrix of the networks. From the normalized weighted adjacency matrix, we computed the correlation of every pair of languages. The correlation analysis between the language pairs can indicate how similar they are, based on how strongly they are connected to the rest of the languages in the network. We observed that English gives a high negative correlation with most of the languages. This is true because English is a very important language; it connects to most of the languages in the network and therefore its similarity with others is weak, given that no other language has a similar connectivity pattern. We then followed with an analysis of correlations without considering English. We represented the correlation matrix diagrams without English. Figure 5.7 and Figure 5.8 represent the correlation matrix of the FLN and the RLN respectively. Here, we discuss a few language relations that interest us. In the FLN,
Swiss German (gsw), Dutch (nl), Hindi (hi), and Urdu (ur) show a high correlation. As we know, Hindi as well as Urdu originate from Sanskrit; Dutch as well as Swiss German originate from Germanic languages and both Germanic and Sanskrit belong to the same family tree, which may be used to explain them being grouped together. We also found Lithuanian (lt), Serbian (sr), Basque (eu), Catalan (ca), and Galician (gl) form a distinct cluster. Except for Basque, all the other languages in the cluster are part of the Indo-European family (although the sub-families are different). Latvian (lv) and Ukranian (uk) are not only part of the same Balto-Slavic sub-family, but also are widely used in Russia. Azerbaijani (az) is a part of the same cluster formed by Latvian and Ukranian, we assume it may be because of the diverse population of countries like Russia¹. Interestingly, although Vietnamese (vi) and Chinese (zh) are not part of the same family, they show similarity probably because they are widely used by natives in China. Similarly, although linguistically very distant, Hungarian (hu) and Danish (da) show similarity because they are widely used in Germany². In the RLN, Bulgarian (bg), Croatian (hr), and German (de) form a cluster. Bulgarian and Croatian are Slavic languages and are popularly used

²http://languageknowledge.eu/countries/germany
in Germany along with German (de). Germanic languages such as Swedish (sv) and Dutch (nl) form a distinct cluster. Although not in the same family tree, Chinese (zh) and Korean (ko) show high correlation; we believe it is because they are widely used in parts of Asia. We also observe that Latin languages like Catalan, Spanish, and Galician form a cluster, and Basque, Spanish, and Portuguese form another cluster. Basque is used in many parts of Spain and some parts of France, which may be the reason why the Basque users interact in the same way as other Latin users in the cluster. Interestingly, the languages in the cluster formed by Indonesian (id), Japanese (ja), Thai (th), Filipino (fil), Vietnamese (vi), Malay (msa), and Hindi (hi) do not have any historic relation, yet the users have a similar interacting pattern, most likely because the languages in the cluster are used by people in neighboring countries in Asia. Likewise, Hungarian (hu), Arabic (ar), and Greek (el) form a cluster although they are linguistically distant. One reason may be because our data was collected during the refugee crisis, which involved Hungary, Syria (Arabic speakers), and Greece. Hence, there were many users who probably put in extra effort to understand the content of the tweets in the above languages. Another interesting cluster is formed by Finnish (fi), Danish (da), Persian (fa), and Bokmal (nb). We believe history plays a role in the
formation of the aforementioned cluster because Finland is located on the border of two regions using languages of two different families: Indo-European (used in Swedish, Danish, Bokmal) and Uralic (used in Finnish, Sami). The languages of the Sami people are now used in the northern parts of Sweden, Finland, Norway, and Russia.

In the analyses above, we attempted to describe why the languages are grouped together; however, it has to be said that this is not an important issue in our work. The explanation for the reason should be left mostly to linguists and historians. What we want to demonstrate is that a Network Science approach may lead to groupings of languages that are a lot more dynamic than historical reasons or family trees. For instance, we argued that the grouping of Syrian, Greek, and Arabic (among others) in the RLN could be due to the current refugee crisis involving these countries. Such indications lead us to believe that the language networks built from online social networks may be a proxy of current events as well as the obvious relations that tend to exist between languages of the same family. However, the connection between such dynamic relations among languages on social networks and the events need further research and investigation. The main purpose of this study was not to claim that we can unveil historical relations between languages or even similarities between languages. Although these could play a role, the contribution of this chapter is to show that we can extract the languages used on Twitter and group them based on population preference, thus revealing a hidden structure of language connectivity in social media. Additionally, the relationships revealed in our work are a more accurate proxy for how information should spread in Twitter, although it is necessary to study the language relations during an event to understand the flow during a phenomenon. The language connectivity is an enabler but other factors should influence the spread. The similarity of languages could be useful to scientists who aim to understand the evolution of languages. Our analyses could be extended with longer periods of Twitter data to have a closer look at the world language relations. That is, the relationships for larger datasets could better reflect long-term relations between languages. Also, bigger datasets can help us look into the temporal aspect of the diffusion of information and how languages can act as a benefit or a barrier to such diffusion. On the other hand, datasets such as ours are more susceptible to variations but are also useful to see how populations interact. In the future, we intend to collect larger datasets and verify if the
groups of languages formed tend to get closer to what linguists or historians expect (language family trees).
Chapter 6

Language as a Measure of Welfare

Our next task was to look at language relations as a function of globalization in society. Globalization is a process driven by international trade that leads to interactions as well as integration of people and government of different nations. Such a process has been impacting us in many ways. Higher economic growth rates and greater affluence are conducive to social development. However, many critics argue that economic growth as measured by GDP cannot capture a nation’s entire development \[61\]. They claim that GDP is intended to measure the productivity of a nation and hence is an insufficient measure to quantify the social development. The social development of a nation or the quality of life (QOL) is a multi-dimensional concept that includes economic growth. The United Nations (UN) introduced a measure called Human Development Index (HDI), which 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. An interesting example is Cuba, which according to the 2007 data, has a low GDP of (PPP) US$6,876 and a very high HDI of 0.863 (the max is 1) \[62\]. The HDI allows countries to be ranked in four tiers: low, medium, high, and very high. The existence of the dominant language in today’s world is also associated with (cultural) globalization. Integration between people from different parts of the world leads to a stronger diffusion of languages, ideas, and values. More recently, the integration has received a further boost by the emergence of online social networks like Twitter because they (online social networks) give us a platform to connect without any geographic restrictions. In fact, they are the quintessential example of globalization.
Yet the links between language usage in society and QOL have 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 chapter looks at how these two concepts (language and QOL) are intertwined; we are interested in understanding the extent to which language usage can be seen as a proxy of QOL.

6.1 Tweet Collection and Network Generation

Although the self-declared language on the profiles of the Twitter users indicates that they have a preferred language to interact in Twitter, sometimes users may tweet in different languages. Keeping this in mind, in this chapter, we collected the languages of the tweets of a user instead of only the declared language. Moreover, such an approach also helps us in an understanding of the language relations at a more granular level. Furthermore, we did not restrict our data collection to any region because region-specific tweets might be biased on languages. We only consider our G20 data among the two datasets we previously investigated (the Follower dataset and the G20 dataset). Details about the G20 data is available in Chapter 3. We also studied another dataset that consists of tweets of the Olympic Games 2016. It is to be noted that we described our additional dataset and the network generation methodology in this section instead of in Chapter 3 because we use the Olympic data as a support for our findings from the G20 data. In addition to considering the tweet language, we also used a different approach to generate our language networks. Hence, our demonstration of the data collection and the network generation in this chapter is aimed for a clear understanding of the methodology used. Below, we describe our data that we used to generate language networks. Table 6.1 shows a few descriptive statistics of the datasets we used.

6.1.1 Additional Dataset: Olympic Games 2016

We used an additional dataset for our analysis in this chapter to confirm our findings. Since the Olympics 2016 was a global event with more than 11,000 athletes from 207 countries, we collected 18 million tweets using the keyword *olympics* in 107 languages. The keyword was
Table 6.1: Descriptive statistics about the datasets. Monolingual refers to users who use one
language and multilingual refers to users who use more than one language.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>G20</th>
<th>Olympics 2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start Data Collection</td>
<td>August 24, 2014</td>
<td>August 01, 2016</td>
</tr>
<tr>
<td>Finish Data Collection</td>
<td>September 29, 2014</td>
<td>August 24, 2016</td>
</tr>
<tr>
<td>Number of Days</td>
<td>35</td>
<td>24</td>
</tr>
<tr>
<td>Number of Users</td>
<td>2,694,784</td>
<td>6,506,634</td>
</tr>
<tr>
<td>Tweets with Identified Language</td>
<td>10,610,653</td>
<td>18,048,522</td>
</tr>
<tr>
<td>Number of Languages in language networks</td>
<td>55</td>
<td>50</td>
</tr>
</tbody>
</table>

translated using Google Translate, as demonstrated in Figure 6.1. More details about this dataset is presented in Table 6.1.

We used the two datasets to generate two language networks. The first step was to collect the percentage of tweets in every language in each dataset. The language that a user mostly uses to tweet (or the language with the highest percentage of tweets) is referred to as the frequent language of the user. We aggregated all the users who demonstrated a particular language as their frequent language. Next, for all the users who use a particular language as their frequent language, we generated a user matrix consisting of the percentages of their tweets in other languages. From the user matrix, we summed the percentages of tweets in other languages and normalized them by the number of users. This information gives us how often the users who use a particular frequent language interact in other languages. In Table 6.2, we demonstrate a toy example. Say we have five users, A, B, C, D and E. The table shows the percentages of tweets of the users in different languages. We see that users A and E use English as their frequent language, B uses Swahili as his frequent language and, C and D use German as their frequent language. The next step is to aggregate the users who demonstrate a particular language as their frequent language.

Figure 6.1: The term “Olympics” was translated to several languages and used to collect tweets related to the Summer Olympic games in Rio de Janeiro, Brazil.
instance, we aggregate the users who demonstrated English as their frequent language (shown in Table 6.3). We also aggregated the users who demonstrated German and Swahili as their frequent languages, and represent that in Table 6.4 and Table 6.5 respectively. We removed the columns with only zeros.

Table 6.2: A toy example to demonstrate the tweet percentages of the users in different languages.

<table>
<thead>
<tr>
<th>Users</th>
<th>English</th>
<th>Spanish</th>
<th>Swahili</th>
<th>German</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.70</td>
<td>0.00</td>
<td>0.00</td>
<td>0.30</td>
</tr>
<tr>
<td>B</td>
<td>0.20</td>
<td>0.00</td>
<td>0.80</td>
<td>0.00</td>
</tr>
<tr>
<td>C</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.70</td>
</tr>
<tr>
<td>D</td>
<td>0.20</td>
<td>0.00</td>
<td>0.00</td>
<td>0.80</td>
</tr>
<tr>
<td>E</td>
<td>0.80</td>
<td>0.00</td>
<td>0.00</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Table 6.3: We used the data in Table 6.2 to generate the matrix of users who demonstrate English as their frequent language. From the user matrix, we generated the normalized English frequency, which is the sum of the tweet frequency in each language divided by the number of users.

<table>
<thead>
<tr>
<th>English Users</th>
<th>German</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.30</td>
</tr>
<tr>
<td>E</td>
<td>0.20</td>
</tr>
<tr>
<td>Normalized English frequency</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Table 6.4: We used the data in Table 6.2 to generate the matrix of users who demonstrate German as their frequent language. From the user matrix, we generated the normalized German frequency, which is the sum of the tweet frequency in each language divided by the number of users.

<table>
<thead>
<tr>
<th>German Users</th>
<th>Spanish</th>
<th>Swahili</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>D</td>
<td>0.00</td>
<td>0.00</td>
<td>0.20</td>
</tr>
<tr>
<td>Normalized German frequency</td>
<td>0.05</td>
<td>0.05</td>
<td>0.15</td>
</tr>
</tbody>
</table>

We calculated the normalized frequency of all the languages to generate a language network, where every node is a language and every edge connected to the node represents the normalized frequency of the language to other languages. For instance, in our toy example, English, German,
Table 6.5: We used the data in Table 6.2 to generate the matrix of users who demonstrate Swahili as their frequent language. From the user matrix, we generated the normalized Swahili frequency, which is the sum of the tweet frequency in each language divided by the number of users.

<table>
<thead>
<tr>
<th>Swahili Users</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0.20</td>
</tr>
<tr>
<td>Normalized Swahili frequency</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Spanish, and Swahili are four nodes. The edge from German to Spanish has weight 0.05, German to Swahili has weight 0.05, German to English has weight 0.15, Swahili to English has edge weight 0.20, and English to German has edge weight 0.25. The language networks are directed. Figure 6.2 and Figure 6.3 show the language networks generated from the G20 and the Olympic datasets.

In both the G20 and the Olympic language networks, we find that English has the highest in-degree. We observe that most of the languages are from the Indo-European language family. A few languages such as Punjabi, Nepali and Telegu have very low in-degree in both the networks. We listed the top five languages according to the in-degree and eigenvector centralities in the G20 as well as the Olympics datasets in Table 6.6 and Table 6.7 respectively. As we mentioned before, the top languages are in advantageous positions in the networks because they are likely to have more information available to them.
Figure 6.2: Language network of the G20 dataset: the color of a node represents the language family and the size of a node represents the in-degree. The edge width represents the normalized tweet frequency between the languages. The represented language families are: Turkic, Indo-European, Japonic, Tai-Kadai, Austronesian, Uralic, Koreanic, Austrasiatic, Sino-Tibetan, Afroasiatic, Dravidian.
Figure 6.3: Language network of the Olympics dataset: the color of a node represents the language family and the size of the node represents the in-degree. The edge width represents the normalized tweet frequency between the languages. The represented language families are: Turkic, Indo-European, Japonic, Tai-Kadai, Austronesian, Uralic, Koreanic, Austrasiatic, Sino-Tibetan, Afroasiatic, Independent, Kartvelian, Dravidian.
Table 6.6: Top languages in the G20 dataset according to in-degree and eigenvector centrality.

<table>
<thead>
<tr>
<th>Indegree</th>
<th>Eigenvector Centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>English</td>
</tr>
<tr>
<td>Indonesian</td>
<td>Indonesian</td>
</tr>
<tr>
<td>Bosnian</td>
<td>Spanish</td>
</tr>
<tr>
<td>Spanish</td>
<td>French</td>
</tr>
<tr>
<td>Estonian</td>
<td>Turkish</td>
</tr>
</tbody>
</table>

Table 6.7: Top languages in the Olympics dataset according to in-degree and eigenvector centrality.

<table>
<thead>
<tr>
<th>Indegree</th>
<th>Eigenvector Centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>English</td>
</tr>
<tr>
<td>Indonesian</td>
<td>Indonesian</td>
</tr>
<tr>
<td>Finnish</td>
<td>Spanish</td>
</tr>
<tr>
<td>Spanish</td>
<td>Portuguese</td>
</tr>
<tr>
<td>Estonian</td>
<td>Italian</td>
</tr>
</tbody>
</table>

6.2 HDI of Languages

The Human Development Index of a language is the average contribution of a single user of the language towards the World HDI and it is calculated using Equation 6.1. We collected the HDI of every country as reported by the United Nations, the language distribution in every country from the data source made available by Ronen et al. [38], and the percentage of speakers of every country for every language as reported in the World Factbook by the Central Intelligence Agency [63]. We computed the HDI of every language by the weighted average below:

\[
\text{HDI}_\ell = \frac{\sum_c (H_c N_{c\ell})}{N_\ell},
\]

(6.1)

where HDI$_\ell$ is the Human Development Index of language $\ell$, $H_c$ is the Human Development Index of a country $c$, $N_{c\ell}$ is the number of speakers of $\ell$ in country $c$, and $N_\ell$ is the total number of speakers of language $\ell$ in the world. Our calculations are based on the data available from several sources; hence the calculated HDI values are approximate.
6.3 Language Rankings by Users and HDI

6.3.1 Languages by the Number of Users

The language distribution of the users can show us the tweet activities in general. Given that Twitter has no restrictions on how users can tweet (except that the tweets have to be less than 140 characters long), the effects are directly reflected in the activities. Figure 6.4 and Figure 6.5 show the contribution of users to the languages in our datasets. According to Figure 6.4 and Figure 6.5, English has the largest demographic followed by Spanish. Although the ranks of the languages are not same, the top languages are similar in both the datasets. As we mentioned previously, it is important to note that our rankings of languages do not reflect the estimates of the world speakers by languages. There are several factors that may influence the different results, for example the penetration of Twitter in the population. However, the disparities do not hinder extracting interesting insights about the users 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 [5].

![Figure 6.4: Distribution of users per language in the G20 dataset](image)

6.3.2 Languages by HDI

In order to analyze the connection between the languages and the development of the countries where they are used, we used Equation 6.1 to calculate the $\text{HDI}_\ell$ of all the languages that are
Figure 6.5: Distribution of users per language in the Summer Olympics dataset.

Present in our datasets. Table 6.8 shows the HDI of the countries. Norway topped the HDI ranks for 12 consecutive years [64]. Hence, a Norwegian user is expected to have better social development. Table 6.9 shows a few languages with very high and very low HDI. The Norwegian language displays a very high HDI and the Haitian language displays a low HDI.

Table 6.8: The five countries with highest and lowest HDI

<table>
<thead>
<tr>
<th>Country</th>
<th>HDI</th>
<th>Country</th>
<th>HDI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Norway</td>
<td>0.94</td>
<td>Nigeria</td>
<td>0.35</td>
</tr>
<tr>
<td>Australia</td>
<td>0.94</td>
<td>Central African Republic</td>
<td>0.35</td>
</tr>
<tr>
<td>Switzerland</td>
<td>0.93</td>
<td>Eritrea</td>
<td>0.39</td>
</tr>
<tr>
<td>Denmark</td>
<td>0.92</td>
<td>Chad</td>
<td>0.39</td>
</tr>
<tr>
<td>Netherlands</td>
<td>0.92</td>
<td>Burundi</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Table 6.9: The five languages with highest and lowest HDI

<table>
<thead>
<tr>
<th>Language</th>
<th>HDI</th>
<th>Language</th>
<th>HDI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Norwegian</td>
<td>0.94</td>
<td>Haitian</td>
<td>0.48</td>
</tr>
<tr>
<td>German</td>
<td>0.91</td>
<td>Punjabi</td>
<td>0.55</td>
</tr>
<tr>
<td>Dutch</td>
<td>0.91</td>
<td>Nepali</td>
<td>0.55</td>
</tr>
<tr>
<td>Swedish</td>
<td>0.91</td>
<td>Khmer</td>
<td>0.56</td>
</tr>
<tr>
<td>Danish</td>
<td>0.91</td>
<td>Urdu</td>
<td>0.59</td>
</tr>
</tbody>
</table>
6.4 Correlation between Network Properties and HDI

We used the centralities of the nodes (languages) in the networks as a measure of their importance. The network properties we considered for our analysis are: in-degree, out-degree, betweenness, closeness, eigenvector centrality, weighted in-degree and weighted out-degree.

We performed a correlation analysis between the network metrics and the HDI\(_\ell\) of the languages. In both the G20 and the Olympics datasets, we found that the HDI\(_\ell\) of the languages correlates significantly with the in-degree as well as the eigenvector centrality of the languages. Our analysis shows that languages that hold such favored positions in the networks also have high HDI\(_\ell\). The users of such languages tend to have a better life at a level more than by chance.

In Figure 6.6 and Figure 6.7, we demonstrate the correlation between the HDI\(_\ell\) of the languages and their positions in the networks. In the in-degree vs HDI\(_\ell\) analysis of both the datasets, we notice that a few popular languages such as Indonesian and French are much lower than the fitting line. Based on their prominence in the network, they were expected to have better HDI\(_\ell\).

We observe that the standard deviation of the HDI of Indonesian users in different countries varies to a great extent. As a result, the variation influences the overall HDI of the language. French does not have a high standard deviation. The low HDI\(_\ell\) of French could be because of many different factors, for example, 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 used in countries having a 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 always related to the HDI. We summarize the results with statistical significance in Table 6.10. The first two rows in Table 6.10 indicate the ones that are statistically significant.

In this chapter, we set to understand if we can relate the language connection patterns of users on Twitter to social development aspects in the real world, such as HDI. Although it is interesting to extract and analyze the entire Twitter language network, it is a very time consuming and expensive procedure. Hence, we used different sets of data from different time periods. In both of our language networks, the language positions correlate with the HDI. 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$\ell$. It is worth noting that some of the popular languages are used in different parts of the world. We demonstrate that overall the positions of languages correlate significantly with the HDI$\ell$. 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 or she chooses to use.
Figure 6.7: The significant positive correlation between language centralities in the Olympics dataset 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.

Table 6.10: The correlation of the network metrics with HDI in the datasets. The first two metrics (rows) have high correlation and are also statistically significant.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>G20</th>
<th>Olympics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>In-Degree</td>
<td>0.42 (p &lt; 0.01)</td>
<td>0.42 (p &lt; 0.01)</td>
<td></td>
</tr>
<tr>
<td>Eigenvector</td>
<td>0.43 (p &lt; 0.01)</td>
<td>0.46 (p &lt; 0.001)</td>
<td></td>
</tr>
<tr>
<td>Out-Degree</td>
<td>0.96 (p &lt; 0.5)</td>
<td>0.29 (p &lt; 0.03)</td>
<td></td>
</tr>
<tr>
<td>Betweenness</td>
<td>0.07 (p &lt; 0.6)</td>
<td>0.08 (p &lt; 0.55)</td>
<td></td>
</tr>
<tr>
<td>Closeness</td>
<td>0.29 (p &lt; 0.04)</td>
<td>0.20 (p &lt; 0.14)</td>
<td></td>
</tr>
<tr>
<td>Weighted In-Degree</td>
<td>0.09 (p &lt; 0.55)</td>
<td>0.08 (p &lt; 0.57)</td>
<td></td>
</tr>
<tr>
<td>Weighted Out-Degree</td>
<td>0.09 (p &lt; 0.5)</td>
<td>-0.19 (p &lt; 0.14)</td>
<td></td>
</tr>
</tbody>
</table>
Chapter 7

Conclusions and Future Work

In today’s world, social networks play a big role in the spread of information. In addition to the connectivity of the users, the languages of the connected users are also an important aspect of the information spread; this is because we live in a world with a diverse set of languages. Our activity patterns on social networks reveal how we connect from the perspective of languages. A better understanding of the structure of languages can reveal several unknown characteristics of the language relations. Such characteristics are more dynamic and evolve from population preference. Hence the characteristics reflect the true patterns of language connectivities rather than the known historical connections. Twitter is a good platform to understand the characteristics of the languages because of the popularity of Twitter, the number of registered users, as well as the number of available languages.

This dissertation investigates the language relations using Network Science approaches. After a brief introduction in Chapter 1 we discussed several interesting works related to our studies in Chapter 2. In Chapter 3 we demonstrated that the languages of the connected users can also be represented in a network form; this network of languages can reveal many interesting characteristics of the language relations. In Chapter 4 we showed that Twitter users tend to connect to other users who are already popular in Twitter. In addition, the users show a strong preference to connect to other users who use similar languages. This preference was observed in both the following and the retweet activities. However, the preference of languages is stronger while retweeting than following. Moreover, the association patterns of languages vary
considerably in Twitter. Some languages show strong association towards themselves while others
do not. One possible explanation why languages associate more with themselves is because they
are already popular and globally used. They do not need much exposure to other languages
in order to receive information. The languages that do not show strong association towards
themselves need to connect to other languages to receive information in the network. Due to
these variations in the connecting patterns, some languages are more disordered than others. In
Chapter 5 we show that the language networks can be used to understand the overall reachability
and similarity of the languages based on the existence and the weight of connections on Twitter.
The reachability of the languages demonstrates how the position of a language in a network can
affect its ability to send or receive information in the network. The similarity of the languages
investigates how languages can be grouped according to their patterns of connections. We also
measured the similarity using the Jaccard similarity measure. Chapter 6 presents how we used
the network metrics of languages to indicate the social development of the users. The in-degree
and eigenvector centrality of the languages correlate significantly with the HDI of the languages.
In a nutshell, our research aims to understand the language relations based on the population
preferences.

Our findings make a significant contribution towards language networks. Furthermore, our
results can have an impact in areas such as the Global Language System, which draws upon
the world system theory, Linguistic Imperialism, and Sociolinguistics, to name a few. We show
many interesting characteristics of the patterns of use of the languages. We hope to collaborate
with linguists in the future to search for an explanation for some of our findings. It is to be
kept in mind that the conclusions we draw in our work are based on the sample data extracted
from Twitter. We also intend to use other sampling strategies to understand if our findings are
biased due to the data collection strategy. Our study can also be extended with longer periods of
Twitter data to have a closer look at the language relations. The language connections extracted
from larger datasets could reflect long-term relations between the languages. Additionally, we
also intend to analyze the tweets that are geo-tagged to understand how frequently a user uses
a language in his or her tweets that is not the dominant language of his or her current location.
The geo-tagged tweets can be particularly helpful in our language positions vs. HDI analysis.
Furthermore, what still remains intriguing to us is to compare the Twitter and the real-world
language connections. This may shed light on the connecting patterns as a whole and allow us to demonstrate whether Twitter is a proxy for society. Although we did not consider other social networks like Facebook, we believe our findings will have similar implications for other social networks. Our research leads to several open and interesting questions. For instance: “Do the language connecting patterns on Twitter reflect the real-world patterns?” and “What can be the structure of languages in other social networks and would they be similar to Twitter?”
Appendix A

List of Publications

The following are the publications that have been produced during the course of this Ph.D. research.

Works directly related with this dissertation


Works not directly related with this dissertation

Bibliography


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