# The Structure and Implications of the Global Language Network

by

Shahar Ronen

B.S., University of Haifa (2007)
M.A., University of Haifa (2011)



Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning
in partial fulfillment of the requirements for the degree of

Master of Science in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2013



Author ........................................................
Program in Media Arts and Sciences
May 10, 2013

Certified by ....................
César A. Hidalgo
Assistant Professor of Media Arts and Sciences
Thesis Supervisor

Accepted by ........................................................
Patricia Maes
Associate Academic Head, Program in Media Arts and Sciences

# The Structure and Implications of the Global Language Network

by

Shahar Ronen

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning
on May 10, 2013, in partial fulfillment of the
requirements for the degree of
Master of Science in Media Arts and Sciences

## Abstract

Languages vary enormously in global importance because of historical, demographic, political, and technological forces, and there has been much speculation about the current and future status of English as a global language. Yet there has been no rigorous way to define or quantify the relative global influence of languages. I propose that the structure of the network connecting multilingual speakers or translated texts, which I call the *Global Language Network*, provides a concept of language importance that is superior to simple economic or demographic measures. I map three independent global language networks (GLN) from millions of records of online and printed linguistic expressions taken from Wikipedia, Twitter, and UNESCO's database of book translations. I find that the structure of the three GLNs is hierarchically organized around English and a handful of hub languages, which include Spanish, German, French, Russian, Malay, and Portuguese, but not Chinese, Hindi or Arabic. Finally, I validate the measure of a language's centrality in the GLNs by showing that it correlates with measures of the number of illustrious people born in the countries associated with that language. I suggest that other phenomena of a language's present and future influence are systematically related to the structure of the global language networks.

Thesis Supervisor: César A. Hidalgo
Title: Assistant Professor of Media Arts and Sciences

**The Structure and Implications of the Global Language Network**

by

Shahar Ronen

The following people served as readers for this thesis:

Thesis Reader ..........................................................
Ethan Zuckerman
Principal Research Scientist
Media Lab

Thesis Reader ..........................................................
Catherine Havasi
Research Scientist
Media Lab

# Acknowledgments

My advisor, *César A. Hidalgo*, for mentoring me through the long and winding road that is science, and for teaching me how to enjoy the ride without losing focus on the destination. Thanks for having complete faith in our work when I had my doubts and for having doubts when I had too much faith. Your curiosity is inspiring and your enthusiasm is contagious.

*Kevin Hu*, my First Mate. You started as a UROP but quickly became an indispensable aide and a full member of our team. Thanks for rising up to any challenge—from scraping through data crunching to web development—and for helping me stir this ship to a safe harbor just in time. I am glad you will be on board in the next few years as well.

My comrades-in-arms at Macro Connections: *Deepak Jagdish*, *Phil Salesses*, *Alex Simoes*, *Daniel Smilkov*, and *Amy Yu*. Thank you for your advice, patience, and constant willingness to brainstorm new ideas. May our friendship, forged in junk food, black coffee and sleepless nights, endure through time and space.

*Bruno Gonçalves*, for providing the Twitter and Wikipedia datasets for mapping the GLN, and *Charles Murray*, for sharing a digital form of his *Human Accomplishment* tables.

This thesis forms the basis of a paper I co-authored with César, Kevin and Bruno, and with *Alessandro Vespignani* and *Steven Pinker*. Collaborating with researchers of such caliber was the best learning experience a grad student could ask for.

The incredible UROP *Michael Xu*, whose sharp mind and quick fingers were instrumental in implementing the first interactive demo of this project.

*Ethan Zuckerman* and *Catherine Havasi*, my readers, whose thoughtful feedback has encouraged me to explore new paths, and whose fresh and unbiased perspective was a true blessing at crunch time.

My parents *Ronny* and *Raiah* and my siblings *Noga* and *Ziv*, thank you for your unconditional love and support since day one.

Finally, my better half, my wife *Noga*, for always being there for me. You are probably even happier than I am to see this thesis submitted.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

> *"...Behold, the people is one, and they have all one language; and this they begin to do: and now nothing will be restrained from them, which they have imagined to do. Go to, let us go down, and there confound their language, that they may not understand one another's speech."*
>
> – Genesis *11:6-7*

Of the thousands of languages that have ever been spoken only a handful have become influential enough to be considered *global languages*. The scarcity of global languages could explain our fascination with headlines such as "Is English or Mandarin the language of the future?" [48] or "It may be time to brush up on your Mandarin" [41], which have become quite common in the last decade. But what determines whether a language becomes global? How do we measure the influence of a language? And what are the implications of a world in which only a handful of languages are globally influential?

In the past, researchers have used a variety of measures to determine the global influence of a language. These include the number of people who speak it, its geographic distribution, the volume of content generated in the language, and the wealth and power of the nations or empires that use it or have used it in the past [18, 46, 50, 74]. Yet demographic and economic measures are unable to capture an important aspect of the global influence of a language [17]: its ability to connect speakers from different languages.

Understanding the rise of a global language is difficult because the processes that determine whether a language becomes global are diverse and often idiosyncratic. One example is network externalities, such as the former use of French in diplomacy or the use of English in air traffic control, in which the widespread use of a standard language for a specific purpose itself forces people in a certain profession to acquire it, making it even more widespread. Major conquests, such as in the spread of the Roman Empire and colonialism, have also increased the linguistic homogeneity of large territories, albeit in less diplomatic ways. Finally, demic expansions, such as the one underlying the spread of agriculture and its Indo-European speakers in Europe [11], contributed to the diffusion of languages in a more distant past. Consequentially, the geographic distribution of languages can teach us about the prehistoric spread of people across Earth [8] and can provide valuable knowledge about the origins of human civilization.

The proper identification of global languages, and the understanding of the mechanisms that give rise to their formation, have political and cultural implications. Policy makers and political movements may be driven by the conflicting goals of promoting a common language that facilitates global communication on one hand and protecting the local languages that strengthen cultural diversity and ethnic or national pride on the other hand. Important decisions therefore hinge on understanding the nature of global languages and the dynamics that give rise to them. Such decisions include the creation and dissemination of legislation that mandates the use of an official language in education, government and public spaces, the subsidy of news and cultural media in a local language, and the investment in technologies for automatic translation. Individuals and businesses who wish to communicate their ideas to a diverse global audience can also benefit from the identification of global languages, which would allow them to make an informed decision about the languages to which they should translate their work.

Finally, linguistic and cultural fragmentations remain important barriers to intercultural exchange in a world where the costs of long-distance communication are historically low. For instance, in the ten countries with the largest online populations, fewer than 8% of the 50 most visited news sites are non-domestic, and in France, only 2% of web news traffic is directed to non-domestic sites [78].

## 1.1 A multilingual world

In an attempt to overcome linguistic barriers, an increasing number of people learn a second or third language [56, 4, 6]. Since learning a new language takes time and effort, people carefully choose which languages to learn. Usually, these are languages that allow them to improve their means of communication. For example, many study English as a second language because it is the *lingua franca* in business, academia, and popular culture [6, 53, 59, 60]. In Switzerland, native speakers of German study French and native speakers of French study German as part of the country's policy to encourage communication between citizens from different language communities [10, 29, 58]. Immigrants learn the language spoken in their new country. Often times their children immerse so well that they do not speak the native language of the parents or they learn it as a second language to remain connected to their heritage [15, 52, 51].

Learning a new language exposes the learner to the influence of another culture and to ideas and information originating from it [24, 25]. People who learn a new language usually retain their connection with their original culture or language community. Thus they become a bridge between their original community and their new community and facilitate the spread of information and ideas between them.

Translations are another channel through which information and ideas diffuse across cultures. While translations spare the need to learn a new language, they are not arbitrary and reflect a demand. After the fall of communism, translations of books from Western Europe to Eastern Europe and former Soviet Union countries increased by a factor of five. Particularly, there was an increase in translations of influential Western works and books by anti-communist authors, reflecting a desire for knowledge that was forbidden during communist times [1].

## 1.2 Measuring the global importance of a language

Which languages should we learn so we could expose ourselves to as many ideas as possible, and communicate our own ideas to as many people as possible? Despite the importance of global languages, there is no rigorous formulation of the concept of a global language, nor a good way to measure the degree to which a language is global. Previous work measured the importance of languages based on their demographics. A ranking of the influence of languages by their number of primary and secondary speakers, the number of countries where they are spoken, and their economic power placed English first, followed by French and Spanish far behind [74]. Ranking languages by the GDP of the countries in which they are spoken placed English first as well—far ahead of Chinese, Japanese, and Spanish [18].

Languages were also ranked by the share of the information their speakers produce of the total information produced world-wide [40]. Information production was defined in this case as the number of books, journals, films, and web pages published in a language. This ranking places English first, with more information produced than the following languages combined, namely German, Spanish, Chinese and French.

The above rankings, however, lack important considerations. The influence of a language is determined not only by its number of speakers, the economic, political and military power of the countries that speak it, and other aggregate attributes, but also by its connections to other languages. A language community is more likely to spread its ideas if it is spoken by many polyglots and is translated to many languages. For example, while Chinese ranks among the top 10 languages in each of the rankings above, it is still an essentially monoglot language community [65, 74], so most of the information produced in Chinese is accessible only to native speakers of the language. Ideas conceived by speakers of Chinese are therefore less likely to spread to other cultures in comparison to ideas conceived by speakers of polyglots language communities such as Spanish or Portuguese.

Studying translations can provide an insight about the accessibility of information created in one language to speakers of other languages. Past studies measured the influence of a language by its share of world-wide book translations [33, 72]. According to this measure, English holds a *hyper-central* position in the world-system of translations based on the share of books translated from English of all book translated worldwide (40% in 1980, a share that has increased since). French, German and Russian were significantly behind, each being the source of 10% to 12% of world translations. However, the above studies did not check to which languages a language was translated, and therefore provide only a limited insight on the diffusion of ideas between language communities.

## 1.3 Focus on the connections

In this thesis I use network science to develop a metric for measuring the global influence of languages and to define what a global language is. My method formalizes the intuition that certain languages are disproportionately influential because they provide direct and indirect paths of translation among most of the world's other languages. For example, it is easy for an idea conceived by a Spaniard to reach a Londoner through bilingual speakers of English and Spanish. An idea conceived by a citizen of Vietnam, however, might only reach a Mapudungun speaker in south-central Chile through a circuitous path that connects bilingual speakers of Vietnamese and English, English and Spanish, and Spanish and Mapudungun. These multilingual speakers are the links between language communities [13]. They define a network that enables the global diffusion of information and ideas, and allow information to flow without a dedicated lingua franca such as Esperanto. I call it the *Global Language Network.*

The idea of a global language network (GLN), which I introduce in this thesis, is a novel approach for evaluating the importance of a language and for studying language connections and potentially the cross-lingual diffusion of ideas. The GLN maps connections between languages using shared speakers and translations, thus shifting the focus from the aggregate measures of languages—number of speakers, income, information production—to the connections between them. The GLN offers a different perspective than phylogenetic trees that connect languages based on words with a similar etymological origin [28], or semantic networks that connect synonyms or words that co-occur frequently in text [35].

The rest of this thesis is organized as follows. Chapter 2 describes the method and the datasets used to map three global language networks—for Twitter, Wikipedia and book translations. Chapter 3 analyzes the three GLNs and their structural similarities. Chapter 4 demonstrates how the GLNs are used to explain the cultural influence of language communities. Finally, Chapter 5 concludes and suggests paths for future research and applications.

**Supporting online material (SOM) for this thesis is available at** `http://macro.media.mit.edu/projects/gln/som`.

# Chapter 2

# Mapping the Global Language Networks

## 2.1 Methods

Finding connections between language communities is challenging. While surveys like the *Eurobarometer* language survey [20, 21] identify polyglots, the number of respondents and their geographical spread is limited. Fortunately, social networking services, blogs, and other platforms for user-generated content allow us to track expressions to individual users, making it possible to identify *bridge figures* that connect language communities [77]. So far, studies that examined the role of individuals in connecting language communities were restricted to a small number of languages, a small number of users, a small number of topics or all of the above. Notable examples include the mapping and comparison of four language networks on the LiveJournal blog service from links found among 6,000 blogs in Portuguese, Russian, Japanese and Finnish [34], and interactions identified among 100,000 blogs that discussed the Haiti earthquake of 2010 in English, Spanish and Japanese [30].

Studies on a larger scale used geographic proximity as a proxy for trans-lingual connections. These studies suggest to connect languages or cultures, or at least measuring their bilateral interest, through requests for Wikipedia pages in languages other than the language associated with the location of the requester [64], or through tweets in different languages made from the same location [44]. While proximity of location may indicate cultural contact, it does not necessarily indicate language contact. Paris is full of tweeting tourists who get exposed to art, cuisine and other forms of French culture during their visit.

However, most of them do not speak French so they are not directly exposed to information and ideas generated in that language, and will not become bridges between the French language community and their native language communities upon their return home.

Studies that map language connections based on a single dataset can draw only a partial picture. There is no single global language network (GLN) because different sets of speakers share different kinds of information across different sets of languages for different purposes. For example, many people use phones and text messages for instant private communication and post on services like Twitter to quickly communicate short-term messages to the public. Fewer people write books, which aim to capture specific knowledge and preserve it for posterity. Accordingly, I map three different versions of the GLN using data from Twitter, Wikipedia, and UNESCO's *Index Translationum* (IT), an international index of printed book translations [69]. I define the *exposure* $e_{ij}$ of language $i$ to language $j$ in each dataset as the conditional probability $P(i|j)$ of observing a connection between the two languages in the dataset. I calculate the exposure for Twitter and Wikipedia as

$$e_{ij} = \frac{M_{ij}}{N_j} \tag{2.1}$$

where $N_j$ represents the number of users with an observed expression in language $j$, and $M_{ij}$ represents the number of users who express themselves in both languages $i$ and $j$. Note that $N_j \leq \sum_i M_{ij}$, since some speakers are fluent in more than two languages and are counted multiple times in $M_{ij}$. The exposure for the book translations dataset is calculated in a slightly different way, as

$$e_{ij} = \frac{N_{i\rightarrow j}}{N_j} \tag{2.2}$$

where $N_{i\rightarrow j}$ represents the number of translations from language $i$ to language $j$ and $N_j$ represents the number of translations into language $j$. Note that for the translations dataset $N_j = \sum_i N_{i\rightarrow j}$ since each individual translation is counted only once (see Section 2.4 for further details on how IT records translations).

In all three cases I merged mutually intelligible languages. For example, Indonesian and Malaysian were both coded as *Malay*, and the regional dialects of Arabic are all coded as *Arabic*. Further information on language notation and merging of languages can be found in Appendix A.

Finally, I note that the estimated probabilities are not symmetrical ($P(i|j) \neq P(j|i)$), and that these asymmetries are often substantial. For example, the probability of observing a user tweeting in English, given that she was observed to tweet in Filipino is 90% ($e_{eng,fil} = 0.9$), whereas the probability of observing that a user tweets in Filipino given that she has been observed to tweet in English is only 2% ($e_{fil,eng} = 0.02$), so $e_{eng,fil} \gg e_{fil,eng}$. I also note that for Twitter and Wikipedia these asymmetries merely reflect the differences in the observed populations (the denominator of Equation 2.1), while for book translations the asymmetries are more meaningful since translations have an inherent direction (Equation 2.2).

The resulting networks represent patterns of linguistic co-expression not among the entire human population but only among the kinds of speakers and texts that contributed to the respective datasets. The populations are confined to literate speakers, and in turn to a subset of social media users (Twitter), book translators (Index Translationum), and knowledgeable public-minded specialists (Wikipedia). Yet these are characteristics of the elites that drive the cultural, political, technological, and economic processes with which observers of global language patterns are concerned. More generally, the tools and constructs developed here may be used to map language networks for any stratum of speakers, given pairwise data on the overlap of language use among them.

The following sections describe in detail the datasets and processes I used to map each GLN and present visualizations that help understand the relative importance of each language.

## 2.2 Twitter

Twitter (`www.twitter.com`) is a microblogging and online social networking service where users communicate using text messages of up to 140 characters long called *tweets*. As of December 2012, Twitter had over 500 million registered users around the world, tweeting in many different languages. Of these, 200 million users were active every month [55].

Tweets are attributed to their authors and can be used to identify polyglots and the language communities they connect, making Twitter a good source for representing the GLN of tens of millions of people. Registered Twitter accounts make up for 7% of world population, but its demographics may not reflect real-life demographics [9]. For example, Twitter users in the United States are younger and hold more liberal opinions than the general public [49]. Twitter is also blocked in China, so the majority of Chinese speakers cannot access it.

I created the initial dataset from 1,009,054,492 tweets collected between December 6, 2012 and February 13, 2012, through the Twitter *garden hose*, which gives access to 10% of all tweets. I detected the language of each tweet using the Chromium Compact Language Detector (CLD) [42], which was chosen for its wide language support and its relatively accurate detection of short messages [31]. However, any automated language detection is prone to errors [34], all the more so when performed on short, informal texts such as tweets. To reduce the effect of such errors, I applied the following methods.

Firstly, to improve detection, I removed *hashtags* (marks of keywords or topics, which start with a #), URLs, and *@-mentions* (references to usernames, which start with a @). Hashtags, URLs and @-mentions are often written in English or in another Latin script, regardless of the actual language of the tweet, and may mislead the detector.

Secondly, I used only tweets that CLD detected with a high degree of confidence. CLD suggests up to three possible languages for the text detected, and gives each option a score that indicates its certainty of the identification, 1 being the lowest and 100 being the highest. If the top option has a much higher score than the other options, CLD marks the identification as *reliable*. I only used tweets that CLD was able to detect with a certainty over 90% and indicated a reliable detection. The 90% threshold was chosen as the optimal tradeoff between detection accuracy and number of tweets detected, based on a sample of 1 million tweets (see Figure 2.1 A).

Thirdly, as mutually intelligible languages are difficult to distinguish, I merged similar languages. To do so, I converted the two-letter ISO 639-1 language codes [36] produced by CLD to three-letter ISO 639-3 codes [61], and merged them using the ISO 639-3 macrolanguages standard. See Appendix A for further details on merging languages.

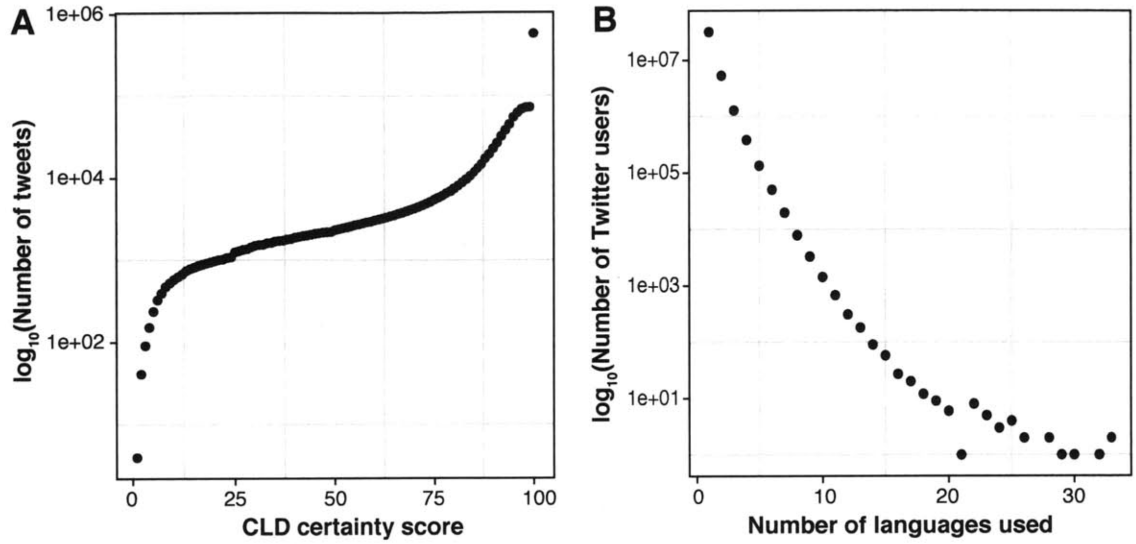


Figure 2.1: **A** Number of tweets as function of certainty **B** Distribution of Twitter users by the number of languages in which they tweet.

Finally, to reduce the effect of individual detection errors, I considered for each user only languages in which he or she tweeted at least twice, and considered only users who made at least five tweets overall. I still found that a large number of users tweeted in a relatively large number of languages, and I attribute some of this to inaccurate language detection. To prevent this from skewing the representation of the Twitter GLN, I discarded users who tweeted in more than five languages (Figure 2.1 B). Five was chosen as the cutoff based on the impression of linguist Richard Hudson that five languages were the most spoken in a community; he coined the term *hyper-polyglots* for people who speak six languages or more [19].[1]

Despite the measures described above, our Twitter dataset still contains detection errors. First, CLD occasionally confuses languages that are similar in their written form but not in their spoken form, such as Urdu and Farsi. Thus, the link between Urdu and Farsi in the Twitter GLN may appear stronger than it actually is. CLD may also confuse languages with no intuitive linguistic connection, such as Japanese and Greek. Japanese tweets often contain *emoji*, Eastern-style emoticons, which may use Greek letters for stylistic purposes,

[1]Some of these users might be bots, which are common on Twitter. Note however that multilingual Twitter bots are not considered a common phenomenon, and even if they were, a bot reading news in one language and re-tweeting them in another is certainly an indication of interaction between the two languages.

such as the kissing emoticon *('ε')* or the crying emoticon *(π_π)*. Japanese tweets that contain emoji may be identified as Greek, especially if they are short enough and contain no (or little) text in addition to the emoji. Thus the link between Japanese and Greek in the Twitter GLN may appear stronger than in reality.

After applying the criteria listed above, I had a dataset of 548,285,896 tweets in 73 languages by 17,694,811 users, who represented over 10% of the active users at the time the data were collected [67]. The clean dataset is available on the SOM page.

I used this dataset to generate the Twitter GLN shown in Figure 2.2. The visualization represents each language as a node. Node sizes are proportional to the number of speakers of each language (native and non-native) as recorded by [76], and node colors indicate language families. Links indicate the strength of the connection between a pair of languages: the color of a link shows the number of users who in tweet in both languages and the width of the link indicates the exposure of one language to another on Twitter. The *exposure* $e_{ij}$ is the conditional probability of a Twitter user to tweet in language $i$ given that he or she tweets in language $j$ (Equation 2.1). For example, for English and Portuguese the dataset lists 10,859,465 users who tweet in English, 1,617,409 who tweet in Portuguese, and 664,320 who tweet in both languages. Therefore, the Twitter exposure of Portuguese to English is 41% ($e_{eng,por} = \frac{664{,}320}{1{,}617{,}409} = 0.41$), whereas the exposure of English to Portuguese is only 6% ($e_{por,eng} = \frac{664{,}320}{10{,}859{,}465} = 0.06$). The Twitter GLN in Figure 2.2 shows only languages that are connected by at least 500 shared Twitter users and have an exposure of at least 0.1% ($e_{ij} \geq 0.001$).

The Twitter GLN consists of 47 nodes and 131 links. Table 2.1 shows statistics for each language (node) in the network. The unfiltered network is available on the SOM page.

## 2.3 Wikipedia

Wikipedia (`www.wikipedia.org`) is a multilingual, web-based, collaboratively edited encyclopedia. As of March 2013, Wikipedia had 40 million registered user accounts across all language editions, of which over 300,000 actively contributed on a monthly basis [43]. Wikipedia's single sign-on mechanism lets editors use the same username on all language

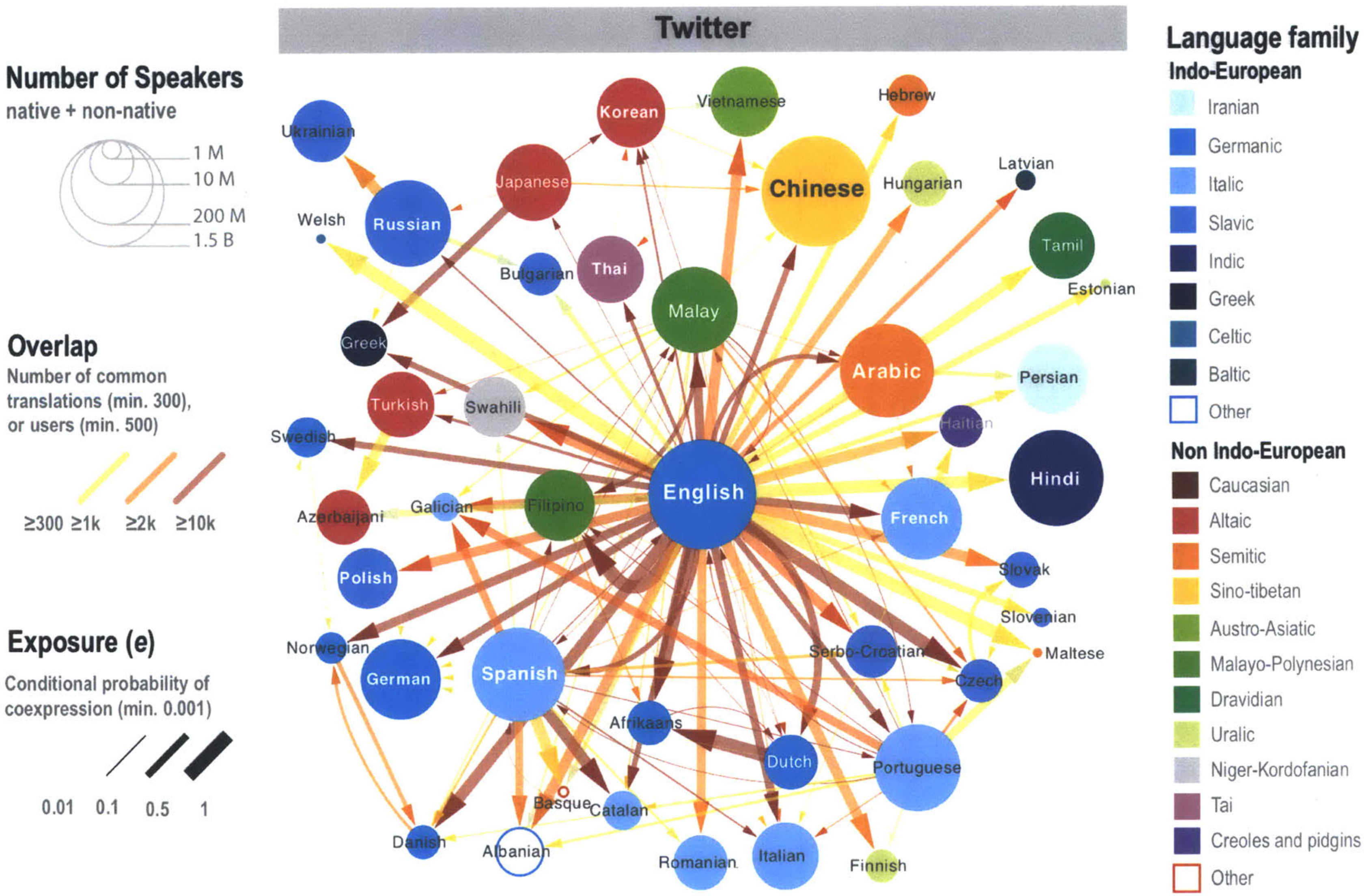


Figure 2.2: The layout of the Twitter global language network. The network contains all the languages that have at least one link whose exposure is 0.1% or more ($e_{ij} \geq 0.001$), with at least 500 shared users.

| | Language | Code | Tweets | Users | Tweets per user | % of total users |
|---|---|---|---|---|---|---|
| 1 | **Afrikaans** | afr | 69,009 | 24,782 | 2.78 | 0.14 |
| 2 | **Albanian** | sqi | 26,682 | 5,155 | 5.18 | 0.03 |
| 3 | **Arabic** | ara | 9,993,172 | 366,643 | 27.26 | 2.07 |
| 4 | **Azerbaijani** | aze | 12,794 | 1,261 | 10.15 | 0.01 |
| 5 | **Basque** | eus | 12,104 | 1,950 | 6.21 | 0.01 |
| 6 | **Bulgarian** | bul | 23,252 | 1,633 | 14.24 | 0.01 |
| 7 | **Catalan** | cat | 236,424 | 32,376 | 7.3 | 0.18 |
| 8 | **Chinese** | zho | 453,837 | 24,113 | 18.82 | 0.14 |
| 9 | **Czech** | ces | 94,324 | 24,573 | 3.84 | 0.14 |
| 10 | **Danish** | dan | 64,537 | 12,029 | 5.37 | 0.07 |
| 11 | **Dutch** | nld | 10,526,980 | 435,128 | 24.19 | 2.46 |
| 12 | **English** | eng | 255,351,176 | 10,859,465 | 23.51 | 61.37 |
| 13 | **Estonian** | est | 22,197 | 2,078 | 10.68 | 0.01 |
| 14 | **Filipino** | fil | 1,905,619 | 257,611 | 7.4 | 1.46 |
| 15 | **Finnish** | fin | 41,165 | 3,856 | 10.68 | 0.02 |
| 16 | **French** | fra | 3,434,065 | 147,843 | 23.23 | 0.84 |
| 17 | **Galician** | glg | 26,035 | 9,302 | 2.8 | 0.05 |
| 18 | **German** | deu | 1,705,256 | 73,897 | 23.08 | 0.42 |
| 19 | **Greek** | ell | 526,527 | 30,609 | 17.2 | 0.17 |
| 20 | **Haitian** | hat | 22,204 | 2,600 | 8.54 | 0.01 |
| 21 | **Hebrew** | heb | 77,937 | 3,384 | 23.03 | 0.02 |
| 22 | **Hindi** | hin | 12,021 | 1,171 | 10.27 | 0.01 |
| 23 | **Hungarian** | hun | 92,093 | 4,804 | 19.17 | 0.03 |
| 24 | **Italian** | ita | 1,586,225 | 89,242 | 17.77 | 0.5 |
| 25 | **Japanese** | jpn | 91,669,691 | 2,602,426 | 35.22 | 14.71 |
| 26 | **Korean** | kor | 11,674,755 | 289,982 | 40.26 | 1.64 |
| 27 | **Latvian** | lav | 168,312 | 13,573 | 12.4 | 0.08 |
| 28 | **Malay** | msa | 49,546,710 | 1,651,705 | 30 | 9.33 |
| 29 | **Maltese** | mlt | 2,838 | 1,156 | 2.46 | 0.01 |
| 30 | **Norwegian** | nor | 170,430 | 16,500 | 10.33 | 0.09 |
| 31 | **Persian** | fas | 79,657 | 2,719 | 29.3 | 0.02 |
| 32 | **Polish** | pol | 167,597 | 8,207 | 20.42 | 0.05 |
| 33 | **Portuguese** | por | 46,520,572 | 1,617,409 | 28.76 | 9.14 |
| 34 | **Romanian** | ron | 73,428 | 5,040 | 14.57 | 0.03 |
| 35 | **Russian** | rus | 4,577,942 | 243,159 | 18.83 | 1.37 |
| 36 | **Serbo-Croatian** | hbs | 54,889 | 8,152 | 6.73 | 0.05 |
| 37 | **Slovak** | slk | 16,657 | 3,657 | 4.55 | 0.02 |
| 38 | **Slovenian** | slv | 21,468 | 2,230 | 9.63 | 0.01 |
| 39 | **Spanish** | spa | 44,195,979 | 2,043,468 | 21.63 | 11.55 |
| 40 | **Swahili** | swa | 32,737 | 5,636 | 5.81 | 0.03 |
| 41 | **Swedish** | swe | 596,130 | 36,604 | 16.29 | 0.21 |
| 42 | **Tamil** | tam | 40,693 | 1,432 | 28.42 | 0.01 |
| 43 | **Thai** | tha | 7,449,790 | 154,171 | 48.32 | 0.87 |
| 44 | **Turkish** | tur | 4,660,694 | 233,158 | 19.99 | 1.32 |
| 45 | **Ukrainian** | ukr | 33,231 | 2,842 | 11.69 | 0.02 |
| 46 | **Vietnamese** | vie | 144,500 | 6,150 | 23.5 | 0.03 |
| 47 | **Welsh** | cym | 5336 | 910 | 5.86 | 0.01 |

Table 2.1: Statistics for languages in the Twitter global language network.

editions to which they contribute. This allows us to associate a contribution with a specific person and identify the languages spoken by that person. Like Twitter, the Wikipedia dataset has its limitations and biases: Wikipedia is blocked in some countries, most notably China, and Wikipedia editors represent neither the general public nor the typical internet user.

I compiled the Wikipedia dataset as follows. Firstly, I used information on editors and their contributions in different languages from the edit logs of all Wikipedia editions until the end of 2011. This information was parsed from Wikipedia's data dumps. I considered only edits to proper articles (as opposed to user pages or talk pages), and only edits made by human editors. Edits by bots used by Wikipedia for basic maintenance tasks (e.g., fixing broken links, spellchecking, adding references to other pages) were ignored, as many of them make changes in an unrealistic number of languages, potentially skewing the GLN. This initial dataset contained 643,435,467 edits in 266 languages by 7,344,390 editors.

Secondly, I merged the languages as I did for the Twitter dataset, discarding ten Wikipedia editions in the process. Two of them are more or less duplicates of other editions, namely

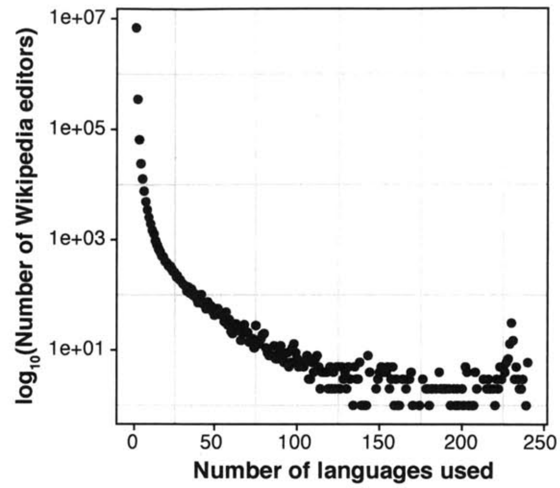


Figure 2.3: Distribution of Wikipedia editors by number of languages in which they contribute.

*simple* (Simple English) of English and *be-x-old* (Classic Belarusian) of Official Belarusian. The remaining eight editions could not be mapped to standard ISO 639-3 languages and were discarded: *bh, cbk_zam, hz, map_bms, nah, nds_nl, tokipona, roa_tara*. These eight editions are small and contain together 220,575 edits by 318 contributors.

Finally, to reduce the effect of one-time edits in given languages editions, which may be cosmetic or technical and may not indicate knowledge of a language, I set the same thresholds as for the Twitter dataset. For each user I considered only languages in which he or she made at least two edits, and considered only users who made at least five edits overall. I also discarded editors who contributed to more than five languages, following the rationale explained in the Twitter section (2.2). I did so because a large number of users contributed to an unrealistic number of languages: hundreds of users contributed to over 50 language editions each, and dozens edited in over 250 languages each (see Figure 2.3). For example, the user Juhko is a self-reported native speaker of Finnish (contributed 6,787 edits to this edition by the end of 2011), and an intermediate speaker of English (834 edits) and Swedish (20). However, Juhko contributed to ten additional language editions, in particular Somali (149 edits) and Japanese (58). Most of these contributions are maintenance work that does not require knowledge of the language, such as the addition of a redirection or the reversion of changes.

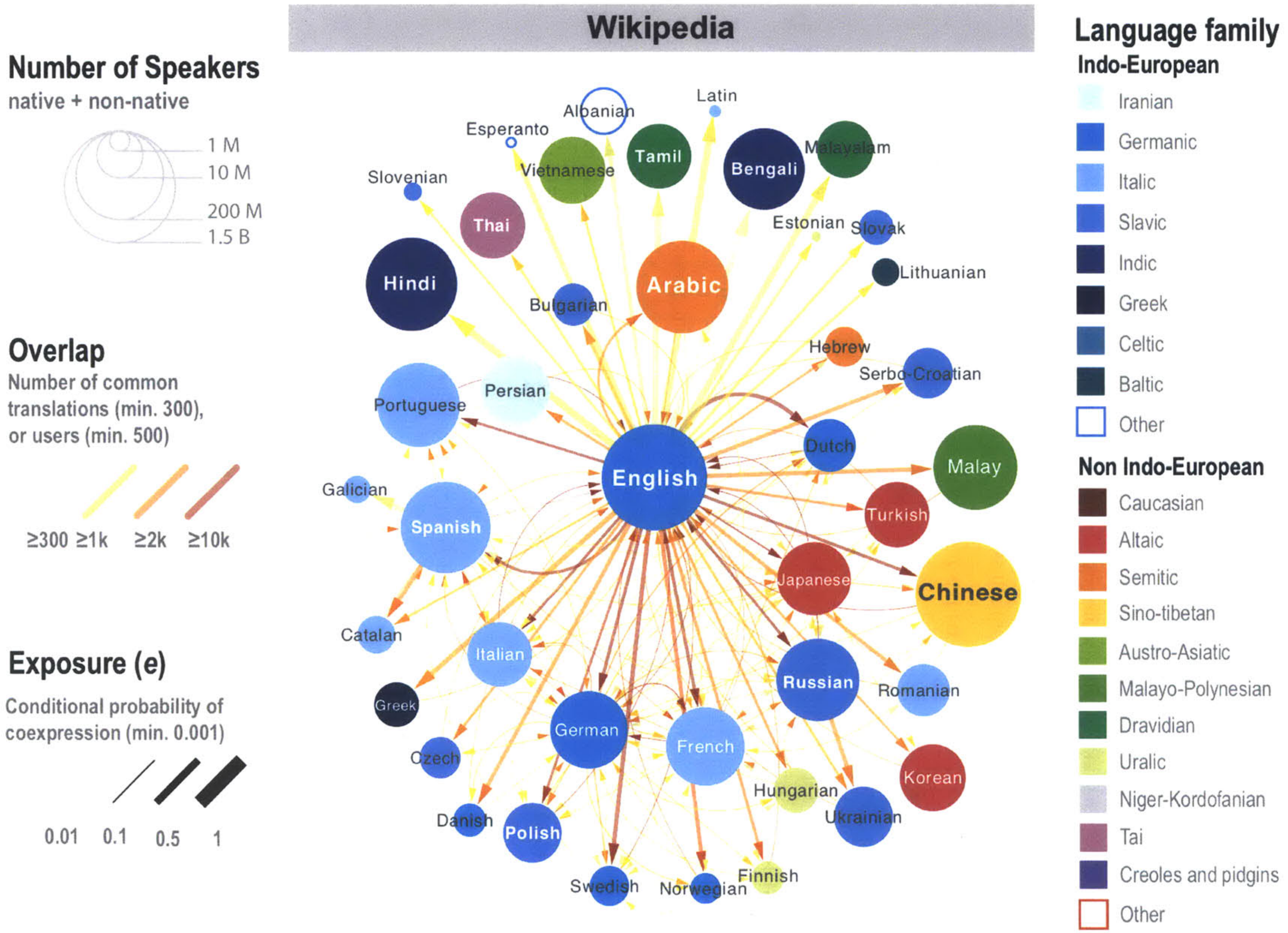


Figure 2.4: The layout of the Wikipedia global language network (GLN). The network contains all the languages that have at least one link whose exposure is 0.1% or more ($e_{ij} \geq 0.001$), with at least 500 shared editors.

| | Language | Code | Edits | Editors | Edits per user | % of total editors |
|---|---|---|---|---|---|---|
| 1 | Albanian | sqi | 196,685 | 1,996 | 98.54 | 0.08 |
| 2 | Arabic | ara | 2,178,719 | 18,258 | 119.33 | 0.71 |
| 3 | Bengali | ben | 147,157 | 1,010 | 145.7 | 0.04 |
| 4 | Bulgarian | bul | 1,130,405 | 6,769 | 167 | 0.26 |
| 5 | Catalan | cat | 1,548,366 | 10,938 | 141.56 | 0.43 |
| 6 | Chinese | zho | 7,302,770 | 50,341 | 145.07 | 1.96 |
| 7 | Czech | ces | 1,697,926 | 15,230 | 111.49 | 0.59 |
| 8 | Danish | dan | 965,082 | 12,270 | 78.65 | 0.48 |
| 9 | Dutch | nld | 6,393,791 | 46,951 | 136.18 | 1.83 |
| 10 | English | eng | 198,361,048 | 1,589,250 | 124.81 | 62.01 |
| 11 | Esperanto | epo | 455,591 | 1,786 | 255.09 | 0.07 |
| 12 | Estonian | est | 366,370 | 3,005 | 121.92 | 0.12 |
| 13 | Finnish | fin | 2,926,115 | 20,811 | 140.6 | 0.81 |
| 14 | French | fra | 23,070,757 | 142,795 | 161.57 | 5.57 |
| 15 | Galician | glg | 246,354 | 1,536 | 160.39 | 0.06 |
| 16 | German | deu | 33,977,378 | 224,215 | 151.54 | 8.75 |
| 17 | Greek | ell | 721,969 | 6,040 | 119.53 | 0.24 |
| 18 | Hebrew | heb | 5,467,149 | 18,998 | 287.77 | 0.74 |
| 19 | Hindi | hin | 310,187 | 1,431 | 216.76 | 0.06 |
| 20 | Hungarian | hun | 2,713,725 | 18,033 | 150.49 | 0.7 |
| 21 | Italian | ita | 11,923,658 | 72,981 | 163.38 | 2.85 |
| 22 | Japanese | jpn | 16,149,315 | 102,857 | 157.01 | 4.01 |
| 23 | Korean | kor | 2,634,092 | 16,464 | 159.99 | 0.64 |
| 24 | Latin | lat | 326,569 | 1,375 | 237.5 | 0.05 |
| 25 | Lithuanian | lit | 363,853 | 3,584 | 101.52 | 0.14 |
| 26 | Malay | msa | 969,369 | 11,005 | 88.08 | 0.43 |
| 27 | Malayalam | mal | 313,554 | 1,435 | 218.5 | 0.06 |
| 28 | Norwegian | nor | 1,789,110 | 22,777 | 78.55 | 0.89 |
| 29 | Persian | fas | 1,603,849 | 14,002 | 114.54 | 0.55 |
| 30 | Polish | pol | 6,589,015 | 47,015 | 140.15 | 1.83 |
| 31 | Portuguese | por | 5,168,734 | 60,487 | 85.45 | 2.36 |
| 32 | Romanian | ron | 852,536 | 11,157 | 76.41 | 0.44 |
| 33 | Russian | rus | 12,445,887 | 81,925 | 151.92 | 3.2 |
| 34 | Serbo-Croatian | hbs | 2,030,039 | 10,901 | 186.23 | 0.43 |
| 35 | Slovak | slk | 433,865 | 4,526 | 95.86 | 0.18 |
| 36 | Slovenian | slv | 456,115 | 5,556 | 82.09 | 0.22 |
| 37 | Spanish | spa | 13,645,596 | 145,487 | 93.79 | 5.68 |
| 38 | Swedish | swe | 3,521,224 | 30,498 | 115.46 | 1.19 |
| 39 | Tamil | tam | 304,589 | 1,289 | 236.3 | 0.05 |
| 40 | Thai | tha | 905,118 | 7,155 | 126.5 | 0.28 |
| 41 | Turkish | tur | 2,062,037 | 23,926 | 86.18 | 0.93 |
| 42 | Ukrainian | ukr | 1,839,988 | 10,028 | 183.49 | 0.39 |
| 43 | Vietnamese | vie | 1,151,775 | 8,244 | 139.71 | 0.32 |

Table 2.2: Statistics for languages in the Wikipedia global language network.

The final dataset consists of 382,884,184 edits in 238 languages by 2,562,860 contributors, and is available on the SOM page. I used this dataset to generate the Wikipedia GLN shown in Figure 2.4, which uses the same visualization conventions used for the Twitter GLN. The visualized network shows only languages that are connected by at least 500 shared Wikipedia editors and have an exposure of at least 0.1% ($e_{ij} \geq 0.001$). For the Wikipedia GLN, the *exposure* $e_{ij}$ is the probability of a Wikipedia editor to contribute to a language edition $i$ given that he or she contributes to language edition $j$ (See Equation 2.1 above). Exposure scores approximate the probability that digitally engaged knowledge specialists speak a pair of languages with a high level of mastery. For example, for German and French, the dataset lists 142,795 editors who contribute to the French Wikipedia edition, 224,215 to the German edition, and 9,236 editors to both. Therefore, the Wikipedia exposure of French to German is 6% ($e_{deu,fra} = \frac{9{,}236}{142{,}795=0.06}$), whereas the exposure of German to French is 4% ($e_{fra,deu} = \frac{9{,}236}{224{,}215} = 0.04$).

Overall, the Wikipedia GLN consists of 43 nodes and 195 links. Table 2.2 shows statistics for each language. The unfiltered network is available on the SOM page.

## 2.4 Book translations

The Index Translationum (IT) is an international bibliography of book translations maintained by UNESCO [69]. The online database contains information on books translated and published in print in about 150 countries since 1979. However, some countries are missing data for certain years, such as translations published in the United Kingdom in the years 1995-2000 and 2009-2011 [68].

IT records translations rather than books, so it does not list books that have not been translated. Moreover, IT also counts each translation separately. For example, IT records 22 independent translations of Tolstoy's *Anna Karenina* from Russian to English. In mapping the network I treat each independent translation separately, and in this case, count 22 translations from Russian to English. Also I note that the source language of a translation recorded by IT can be different from the language in which the book was originally written. For example, the IT records 15 translations of *The Adventures of Tom Sawyer* to Catalan (as of March 2013), but only 13 were translated directly from the original English; the other two are from Spanish and Galician. This characteristic of the dataset allows me to identify languages that serve as intermediaries for translations.

I retrieved a dump of the data on July 22, 2012, which contained 2,244,527 translations in 1,160 languages. After removing a few corrupt entries, I converted the language codes listed in IT to standard three-letter ISO 639-3 codes. The following entries were discarded from the dataset: 41 miscellaneous dialects of languages that were already listed (together accounting for under 100 translations total), 46 languages that could not be mapped to standard ISO 639-3 codes (together accounting for about a thousand translations total), and five administrative codes (*mis, mul, und, zxx*, and *not supplied*; see [61]). The remaining languages were merged into macrolanguages (see Appendix A).

The cleaned dataset contains 2,231,920 translations in 1,019 languages. I used this dataset to generate the book translation GLN shown in Figure 2.5. This network shows languages that are connected by at least 300 translations and have an exposure of at least 0.1% ($e_{ij} \geq 0.001$). The *exposure* $e_{ij}$ is the conditional probability of a book to have been translated from language $i$ given that the book was translated into language $j$ (Equa-

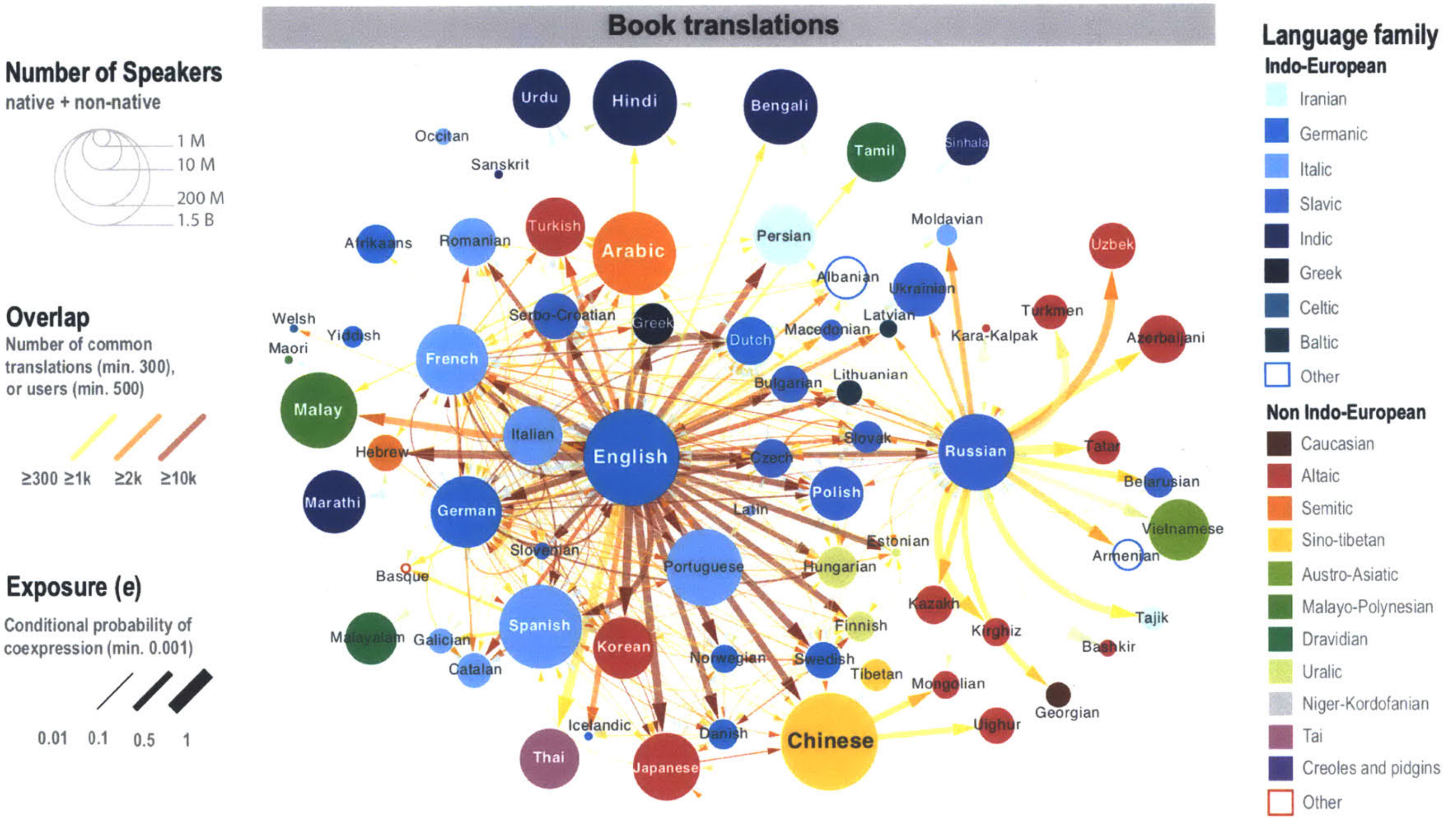


Figure 2.5: The layout of the book translation global language network. The network contains all the languages that have at least one link whose exposure is 0.1% or more ($e_{ij} \geq 0.001$), with at least 300 translations.

tion 2.2). For example, for English and Hebrew our dataset lists 146,294 total translations into English, of which 2,831 translations are from Hebrew. Therefore, the translation exposure of English to Hebrew is 0.2% ($e_{heb,eng} = \frac{2{,}831}{146{,}294} = 0.002$). Because there are 10,961 total translations to Hebrew, of which 8,620 translations are from English, the translation exposure of Hebrew to English is 79% ($e_{eng,heb} = \frac{8{,}620}{10{,}961} = 0.79$).

I removed three languages that met the thresholds for translations and exposure, but are no longer in use: Ancient Greek (ISO 639-3 identifier *grc*), Middle High German (*gmh*), and Old French (*fro*). Overall, the book translation GLN consists of 71 nodes and 500 links. Table 2.3 shows statistics for each language. The unfiltered network is available on the SOM page.

| | Language | Code | Translations from | Translations to |
|---|---|---|---|---|
| 1 | **Afrikaans** | afr | 357 | 776 |
| 2 | **Albanian** | sqi | 1,424 | 6,757 |
| 3 | **Arabic** | ara | 11,884 | 12,488 |
| 4 | **Armenian** | hye | 1,100 | 2,139 |
| 5 | **Azerbaijani** | aze | 774 | 1,658 |
| 6 | **Bashkir** | bak | 357 | 502 |
| 7 | **Basque** | eus | 1,021 | 3,923 |
| 8 | **Belarusian** | bel | 1,409 | 1,874 |
| 9 | **Bengali** | ben | 2,223 | 1,878 |
| 10 | **Bulgarian** | bul | 3,667 | 25,742 |
| 11 | **Catalan** | cat | 8,328 | 18,004 |
| 12 | **Chinese** | zho | 13,337 | 62,650 |
| 13 | **Czech** | ces | 17,202 | 64,442 |
| 14 | **Danish** | dan | 21,239 | 64,799 |
| 15 | **Dutch** | nld | 18,978 | 111,371 |
| 16 | **English** | eng | 1,225,237 | 146,294 |
| 17 | **Estonian** | est | 5,739 | 20,605 |
| 18 | **Finnish** | fin | 8,296 | 46,271 |
| 19 | **French** | fra | 216,624 | 238,463 |
| 20 | **Galician** | glg | 1,346 | 2,371 |
| 21 | **Georgian** | kat | 1,224 | 2,176 |
| 22 | **German** | deu | 201,718 | 292,124 |
| 23 | **Greek** | ell | 4,862 | 27,422 |
| 24 | **Hebrew** | heb | 9,889 | 10,961 |
| 25 | **Hindi** | hin | 1,469 | 3,506 |
| 26 | **Hungarian** | hun | 11,256 | 54,989 |
| 27 | **Icelandic** | isl | 1,518 | 6,514 |
| 28 | **Italian** | ita | 66,453 | 59,830 |
| 29 | **Japanese** | jpn | 26,921 | 130,893 |
| 30 | **Kara-Kalpak** | kaa | 129 | 568 |
| 31 | **Kazakh** | kaz | 948 | 2,454 |
| 32 | **Kirghiz** | kir | 708 | 1,528 |
| 33 | **Korean** | kor | 4,621 | 22,338 |
| 34 | **Latin** | lat | 19,240 | 362 |
| 35 | **Latvian** | lav | 1,288 | 8,145 |
| 36 | **Lithuanian** | lit | 1,985 | 15,447 |

| | Language | Code | Translations from | Translations to |
|---|---|---|---|---|
| 37 | **Macedonian** | mkd | 1,592 | 3,901 |
| 38 | **Malay** | msa | 485 | 5,416 |
| 39 | **Malayalam** | mal | 306 | 1,202 |
| 40 | **Maori** | mri | 88 | 319 |
| 41 | **Marathi** | mar | 405 | 878 |
| 42 | **Moldavian** | mol | 2,864 | 3,720 |
| 43 | **Mongolian** | mon | 244 | 2,423 |
| 44 | **Norwegian** | nor | 14,530 | 45,923 |
| 45 | **Occitan** | oci | 452 | 204 |
| 46 | **Persian** | fas | 2,837 | 11,329 |
| 47 | **Polish** | pol | 14,104 | 76,720 |
| 48 | **Portuguese** | por | 11,390 | 74,721 |
| 49 | **Romanian** | ron | 5,475 | 18,464 |
| 50 | **Russian** | rus | 101,395 | 82,772 |
| 51 | **Sanskrit** | san | 4,282 | 58 |
| 52 | **Serbo-Croatian** | hbs | 12,743 | 45,036 |
| 53 | **Sinhala** | sin | 52 | 671 |
| 54 | **Slovak** | slk | 4,205 | 19,641 |
| 55 | **Slovenian** | slv | 2,463 | 18,719 |
| 56 | **Spanish** | spa | 52,955 | 228,910 |
| 57 | **Swedish** | swe | 39,192 | 71,688 |
| 58 | **Tajik** | tgk | 476 | 1,062 |
| 59 | **Tamil** | tam | 496 | 1,763 |
| 60 | **Tatar** | tat | 462 | 819 |
| 61 | **Thai** | tha | 215 | 1,227 |
| 62 | **Tibetan** | bod | 1,508 | 344 |
| 63 | **Turkish** | tur | 2,658 | 11,874 |
| 64 | **Turkmen** | tuk | 434 | 741 |
| 65 | **Uighur** | uig | 81 | 1,488 |
| 66 | **Ukrainian** | ukr | 2,877 | 4,514 |
| 67 | **Urdu** | urd | 950 | 1,005 |
| 68 | **Uzbek** | uzb | 872 | 2,757 |
| 69 | **Vietnamese** | vie | 668 | 786 |
| 70 | **Welsh** | cym | 621 | 2,312 |
| 71 | **Yiddish** | yid | 1,590 | 89 |

Table 2.3: Statistics for languages in the books translation global language network.

# Chapter 3

# Analysis

The three GLNs presented in Figures 2.2, 2.4 and 2.5 share a number of features. First, the number of expressions observed in each language—Twitter users, Wikipedia editors, or translations from a language—correlates strongly across the three networks (Figures 3.1 A-C). Moreover, the exposures of the multilingual links correlate strongly across the three networks (Figures 3.1 D-F), in particular Twitter-Wikipedia and Wikipedia-book translations. This means that a language with a high or low exposure to another language in one network is likely to have a similar exposure to the same language in the other networks.

## 3.1 Degree distribution

The three networks also share several structural features. First, the three GLNs exhibit a *scale-free structure* [5]. Let the *connectivity* or *degree* $k_i$ of a language $i$ be the number of other languages connected to it. All three networks have *long-tailed* degree distributions, and their cumulative probability distributions are well approximated by the *power law* behavior $P(k \geq k^*) \sim k^{-2}$ for $k^* > 5$ (Figures 3.2 A-C). That is, the probability of a language to have a degree $k^*$ or larger decreases following the above power law as $k^*$ increases. This behavior highlights the disproportionately high degree of hub languages. Only two of the 47 languages in the Twitter GLN (English and Malay) are connected to 20 other languages or more, and only two of the 43 languages in the Wikipedia GLN (English

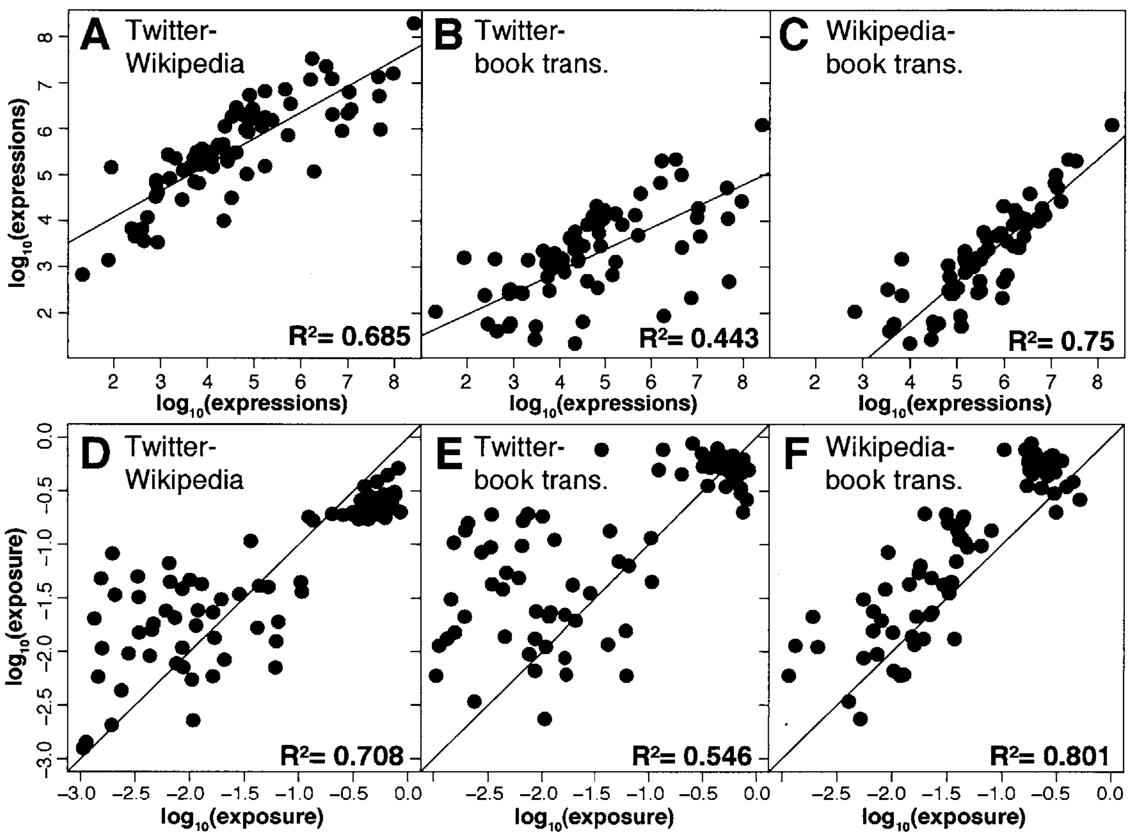


Figure 3.1: Similarity of the three independent datasets I use for mapping the global language networks. The top row shows the correlation between the number of expressions across the three datasets: **A** tweets and Wikipedia edits in a language **B** tweets in a language and translations from a language **C** Wikipedia edits in a language and book translations from a language. The bottom row shows the correlation between the exposures ($e$) measured for language pairs in the **D** Twitter and Wikipedia GLNs, **E** Twitter and book translation GLNs, and **F** Wikipedia and book translations GLNs.

and German). In the book translation GLN, only six languages of 71 (English, Russian, French, German, Spanish and Italian) are connected to more than 20 languages.

## 3.2 Clustering-connectivity

Moreover, the three GLNs share what is known as a *hierarchical structure* [62]. A network is considered to be hierarchical if the more connected its nodes are, the less likely their neighbors are to be a clique. The method I use to measure the hierarchical structure of each GLN was adapted from a method used to measure hierarchy in protein-interaction and technological networks [54, 71].

The probability that the neighbors of a node are connected to each other is expressed by the node's *local clustering coefficient* [73]. Formally, the clustering coefficient $C_i$ of language $i$ is defined as $C_i = \frac{2\Delta_i}{k_i(k_i-1)}$, where $k_i$ is the degree of the language, $\Delta_i$ is the observed number of fully-connected triplets (3-cliques) for the neighbors of $i$, and $\frac{k_i(k_i-1)}{2}$ is the number of possible fully connected triplets for the neighbors of $i$ (the number of ways of choosing two nodes from the $k_i$ neighbors of language $i$). In both cases I count triplets in an undirected version of the network. Then, I plot the clustering coefficient $C_i$ of each node $i$ as a function of its degree $k_i$. In a hierarchical network, the clustering of a node will be inversely related to its degree [54].

The hierarchical structure of the GLNs is illustrated in Figures 3.2 D-F. The hierarchy is characterized by an exponential decay of clustering as a function of connectivity, which is faster than the power-law decay observed in biological and technological networks [54, 71]. In the GLN, the inverse relationship between clustering and connectivity means that hub languages are linked to clusters of languages that are connected within themselves but are not directly connected to languages in other clusters. Hence, the hierarchical structure of the GLN indicates that hub languages act as bridges between languages from different clusters. English is the major hub in all three GLNs. The intermediate hubs include Malay, Spanish and Portuguese in the Twitter network, German in the Wikipedia network, and Russian, French and German in the book translation network. These findings agree with previous studies examining book translations, which concluded that English held a

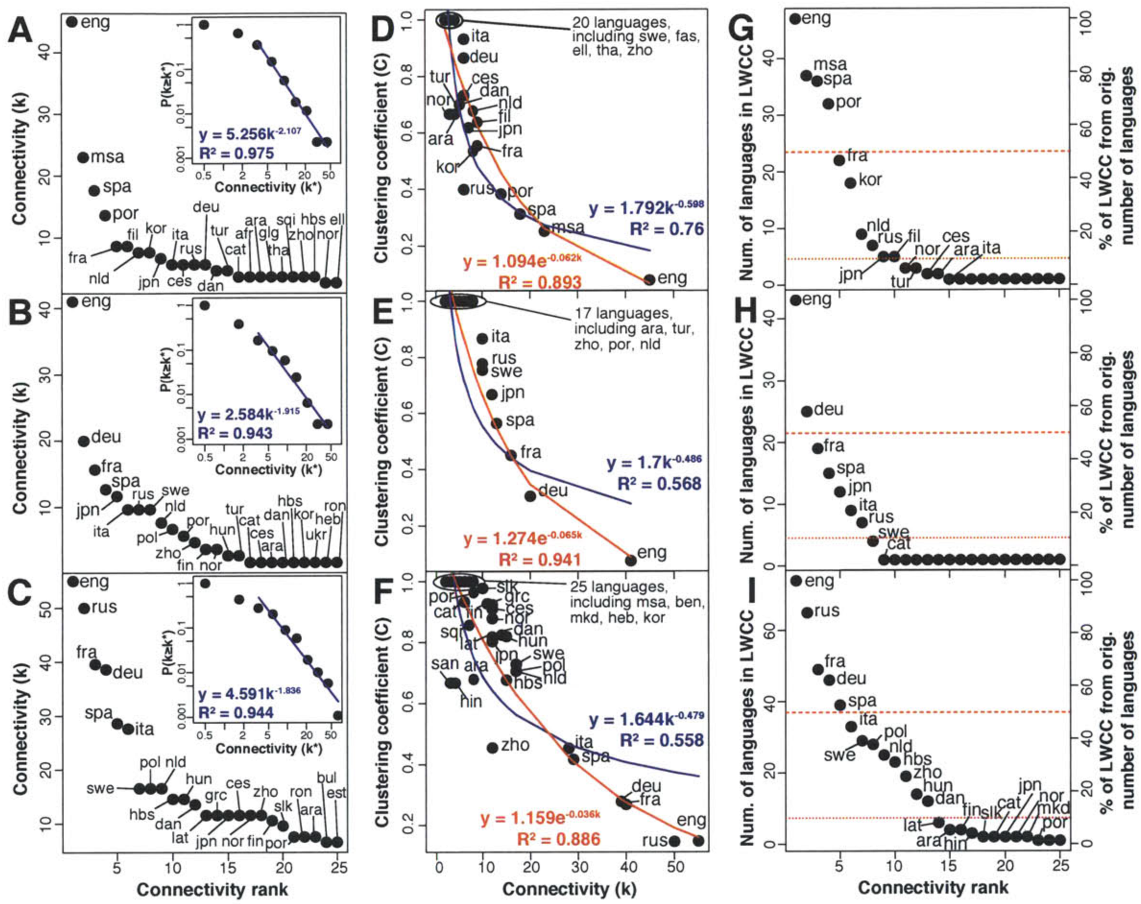

Figure 3.2: Analysis of the structure of the global language networks. Degree ranking diagrams, with cumulative degree distributions in the inset, for **A** Twitter GLN, **B** Wikipedia GLN, and **C** book translation GLN. Clustering-connectivity diagrams, showing the clustering of each language as a function of its connectivity: for **D** Twitter GLN, **E** Wikipedia GLN, and **F** book translation GLN. Percolation analysis diagrams, showing the size of the largest weakly connected component (LWCC) upon removing the $n^{th}$ most connected language (connectivity is re-calculated after removing each node): for **G** Twitter GLN, **H** Wikipedia GLN, and **I** book translation GLN. The top horizontal line marks the 50% threshold, and the dashed line marks 10%.

*hyper-central* position in the world-system of translations, followed by French, German and Russian [33, 72].

The hierarchical structure of the networks means that the paths connecting peripheral languages go first through nodes in increasing order of connectivity, and then through nodes in decreasing order of connectivity [66]. For example, in the book translation GLN the path between Kazakh and Galician goes through nodes in increasing order of connectivity from Kazakh to Russian and from Russian to English, and then through nodes in decreasing order of connectivity from English to Spanish and from Spanish to Galician. Here, Kazakh and Galician are peripheral languages in this GLN, Russian and Spanish are intermediate hubs, and English is the main hub.

## 3.3 Percolation analysis

Finally, I explore the implications of the hierarchical structure of the GLN. I do so by measuring the size of the network's *largest weakly-connected component (LWCC)* as nodes are removed from the network in decreasing order of connectivity, a method known as *percolation analysis* [14]. The LWCC of a network is the largest subset of nodes for which there is an undirected path between every pair of nodes. Percolation analyses of this kind have been used to study the vulnerability of networks to errors and attacks: due to their nature, scale-free networks were found to be extremely vulnerable to attacks, that is, the removal of their hubs [3].

Figures 3.2 G-I show that the three GLNs become quickly disconnected when a few hub languages are removed. In all cases, the removal of five hubs or fewer reduced the largest connected component to half its original size. People who do not speak these hub languages are very limited in their ability to communicate with people from most other cultures, and if these languages suddenly vanished off the face of the earth, global communication would become extremely difficult. Removing 14, 8, and 22 languages from the Twitter, Wikipedia, and book translation networks, respectively, reduced the largest connected component in each network to a dyad. In such a situation, global communication would be impossible.

# Chapter 4

# Language centrality and cultural contribution

To demonstrate an application of the GLN, I study the relationship between the position of a language in the GLN and the global cultural influence of its speakers, and compare it with the relationship between the cultural influence of a language and its population and income. I measure the position of a language in the GLN using its *eigenvector centrality* [7]. Eigenvector centrality considers the connectivity of a language as well as that of its neighbors, and that of its neighbors' neighbors, in an iterative manner. Hence, eigenvector centrality rewards hubs that are connected to other hubs (a variant of this method is also the basis for Google's PageRank algorithm [47]). Table 4.1 lists the eigenvector centrality for each language in each of the three GLNs. The sources for the population and income data and their preparation are explained in detail in Appendix B.

## 4.1 Cultural contribution dataets

I measure the cultural impact of a language through the number it speakers that made a long-lasting cultural impression on the world. I focus on these *illustrious people*, rather than on ideas or other forms of cultural expression, because people names are easier to identify and match across languages.

| | Language | Code | Twitter | Wikipedia | Book translations |
|---|---|---|---|---|---|
| 1 | **Afrikaans** | afr | 0.28 | | 0.03 |
| 2 | **Albanian** | sqi | 0.21 | 0.05 | 0.2 |
| 3 | **Arabic** | ara | 0.27 | 0.17 | 0.32 |
| 4 | **Armenian** | hye | | | 0.06 |
| 5 | **Azerbaijani** | aze | 0.09 | | 0.06 |
| 6 | **Bashkir** | bak | | | 0.03 |
| 7 | **Basque** | eus | 0.1 | | 0.16 |
| 8 | **Belarusian** | bel | | | 0.06 |
| 9 | **Bengali** | ben | | 0.05 | 0.1 |
| 10 | **Bulgarian** | bul | 0.08 | 0.1 | 0.29 |
| 11 | **Catalan** | cat | 0.32 | 0.16 | 0.28 |
| 12 | **Chinese** | zho | 0.16 | 0.4 | 0.35 |
| 13 | **Czech** | ces | 0.43 | 0.18 | 0.47 |
| 14 | **Danish** | dan | 0.19 | 0.18 | 0.5 |
| 15 | **Dutch** | nld | 0.48 | 0.58 | 0.54 |
| 16 | **English** | eng | 1 | 1 | 1 |
| 17 | **Esperanto** | epo | | 0.05 | |
| 18 | **Estonian** | est | 0.06 | 0.05 | 0.29 |
| 19 | **Filipino** | fil | 0.58 | | |
| 20 | **Finnish** | fin | 0.09 | 0.32 | 0.41 |
| 21 | **French** | fra | 0.51 | 0.82 | 0.91 |
| 22 | **Galician** | glg | 0.25 | 0.07 | 0.12 |
| 23 | **Georgian** | kat | | | 0.06 |
| 24 | **German** | deu | 0.35 | 0.88 | 0.95 |
| 25 | **Greek** | ell | 0.17 | 0.1 | 0.27 |
| 26 | **Haitian** | hat | 0.11 | | |
| 27 | **Hebrew** | heb | 0.06 | 0.18 | 0.24 |
| 28 | **Hindi** | hin | 0.06 | 0.05 | 0.1 |
| 29 | **Hungarian** | hun | 0.06 | 0.26 | 0.48 |
| 30 | **Icelandic** | isl | | | 0.18 |
| 31 | **Italian** | ita | 0.47 | 0.67 | 0.67 |
| 32 | **Japanese** | jpn | 0.27 | 0.72 | 0.49 |
| 33 | **Kara-Kalpak** | kaa | | | 0.03 |
| 34 | **Kazakh** | kaz | | | 0.07 |
| 35 | **Kirghiz** | kir | | | 0.06 |
| 36 | **Korean** | kor | 0.4 | 0.16 | 0.2 |
| 37 | **Latin** | lat | | 0.05 | 0.26 |
| 38 | **Latvian** | lav | 0.06 | | 0.16 |
| 39 | **Lithuanian** | lit | | 0.05 | 0.23 |
| 40 | **Macedonian** | mkd | | | 0.09 |
| 41 | **Malay** | msa | 0.82 | 0.1 | 0.05 |
| 42 | **Malayalam** | mal | | 0.05 | 0.03 |
| 43 | **Maltese** | mlt | 0.09 | | |
| 44 | **Maori** | mri | | | 0.03 |
| 45 | **Marathi** | mar | | | 0.03 |
| 46 | **Moldavian** | mol | | | 0.13 |
| 47 | **Mongolian** | mon | | | 0.04 |
| 48 | **Norwegian** | nor | 0.09 | 0.32 | 0.45 |
| 49 | **Occitan** | oci | | | 0.03 |
| 50 | **Persian** | fas | 0.09 | 0.1 | 0.2 |
| 51 | **Polish** | pol | 0.06 | 0.52 | 0.55 |
| 52 | **Portuguese** | por | 0.57 | 0.46 | 0.35 |
| 53 | **Romanian** | ron | 0.1 | 0.18 | 0.34 |
| 54 | **Russian** | rus | 0.22 | 0.64 | 0.86 |
| 55 | **Sanskrit** | san | | | 0.07 |
| 56 | **Serbo-Croatian** | hbs | 0.17 | 0.18 | 0.45 |
| 57 | **Sinhala** | sin | | | 0.03 |
| 58 | **Slovak** | slk | 0.11 | 0.05 | 0.36 |
| 59 | **Slovenian** | slv | 0.06 | 0.05 | 0.24 |
| 60 | **Spanish** | spa | 0.69 | 0.72 | 0.78 |
| 61 | **Swahili** | swa | 0.14 | | |
| 62 | **Swedish** | swe | 0.12 | 0.64 | 0.57 |
| 63 | **Tajik** | tgk | | | 0.03 |
| 64 | **Tamil** | tam | 0.06 | 0.05 | 0.06 |
| 65 | **Tatar** | tat | | | 0.06 |
| 66 | **Thai** | tha | 0.22 | 0.05 | 0.03 |
| 67 | **Tibetan** | bod | | | 0.05 |
| 68 | **Turkish** | tur | 0.31 | 0.26 | 0.17 |
| 69 | **Turkmen** | tuk | | | 0.03 |
| 70 | **Uighur** | uig | | | 0.04 |
| 71 | **Ukrainian** | ukr | 0.03 | 0.16 | 0.13 |
| 72 | **Urdu** | urd | | | 0.07 |
| 73 | **Uzbek** | uzb | | | 0.06 |
| 74 | **Vietnamese** | vie | 0.1 | 0.05 | 0.03 |
| 75 | **Welsh** | cym | 0.06 | | 0.07 |
| 76 | **Yiddish** | yid | | | 0.07 |

Table 4.1: Eigenvector centrality by language in each of the three GLNs (rounded to the nearest hundredth).

I use two separate methods to decide whether a person is illustrious. The first is having Wikipedia articles in at least 20 language editions, and the second is being included in the *Human Accomplishment* list [45], a list of 3,869 influential people in the arts and sciences, from 800 BCE to 1950. As neither dataset contains information about the language used by the illustrious people it lists, I start this section by describing how I associated illustrious people with languages. Then, I dedicate a subsection to each dataset, in which I describe how the dataset was retrieved and prepared for use.

### 4.1.1 Associating illustrious people with languages

Ideally each language would be given a point for each notable person who spoke this language as his or her native language, or who used this language as the main language for his or her main contributions. Unfortunately, this information is not available in a structured format and finding it manually for each person does not scale well for thousands of people. Therefore, I determined a person's language affiliation using the current language demographics for his or her country of birth. Each illustrious person in the datasets equals one point, which is distributed across the languages spoken in his or her native country according to their population [38, 12]. For example, Italian inventor Guglielmo Marconi counts as one point for Italian. Former Canadian Prime Minister Pierre Trudeau contributes 0.65 to English, 0.35 to French. I stress again that my scoring is based on national identity and not on cultural or linguistic identity. Trudeau was a native speaker of French while Leonard Cohen is a native speaker of English, but since both of them are Canadian, each one adds 0.65 points for English and 0.35 points for French, regardless of their native language. Refer to Table 4.2 for the language demographics of each country.

I determine a person's country of birth using present-day international borders. For example, I code Italy as the country of birth for author Ippolito Nievo, although Italy was unified only shortly before his death in 1861 and at the time of his birth his native Padua was part of the Austrian Empire. This method produces unintuitive results: the Ancient Greek historian Herodotus was born in Halicarnassus (present-day Bodrum, Turkey) and would earn points for Turkish, while Mustafa Kemal Atatürk, founder of the Republic of Turkey,

| | Country | Languages |
|---|---|---|
| 1 | Afghanistan | Persian: 50% |
| 2 | Albania | Albanian: 95%, Romanian: 5% |
| 3 | Algeria | Arabic: 73%, French: 5% |
| 4 | American Samoa | English: 2.9% |
| 5 | Andorra | Catalan: 50%, Spanish: 40%, French: 10% |
| 6 | Angola | Portuguese: 80% |
| 7 | Anguilla | English: 100% |
| 8 | Antigua and Barbuda | English: 100% |
| 9 | Argentina | Spanish: 85%, Italian: 3.8% |
| 10 | Armenia | Armenian: 97.7%, Russian: 0.9% |
| 11 | Aruba | Spanish: 12.6%, English: 7.7%, Dutch: 5.8% |
| 12 | Australia | English: 78.5%, Chinese: 2.5%, Italian: 1.6%, Greek: 1.3%, Arabic: 1.2% |
| 13 | Austria | German: 88.6%, Turkish: 2.3%, Serbo-Croatian: 2.2% |
| 14 | Azerbaijan | Azerbaijani: 90.3%, Russian: 1.8%, Armenian: 1.5% |
| 15 | Bahamas | English: 100% |
| 16 | Bahrain | Arabic: 100% |
| 17 | Bangladesh | Bengali: 98%, English: 2% |
| 18 | Barbados | English: 100% |
| 19 | Belarus | Russian: 62.8%, Belarusian: 36.7% |
| 20 | Belgium | Dutch: 60%, French: 40% |
| 21 | Belize | Spanish: 46%, Haitian: 32.9%, English: 3.9% |
| 22 | Benin | French: 40% |
| 23 | Bermuda | English: 91.8%, Portuguese: 4% |
| 24 | Bolivia | Spanish: 60.7%, Quechua: 21.2%, Aymara: 14.6% |
| 25 | Bosnia and Herzegovina | Serbo-Croatian: 33.2% |
| 26 | Botswana | English: 2.1% |
| 27 | Brazil | Portuguese: 100% |
| 28 | Brunei Darussalam | Malay: 80%, English: 20% |
| 29 | Bulgaria | Bulgarian: 84.5%, Turkish: 9.6% |
| 30 | Burkina Faso | French: 100% |
| 31 | Burundi | French: 0.02% |
| 32 | Cambodia | Central Khmer: 95%, French: 2.5%, English: 2.5% |
| 33 | Cameroon | French: 50%, English: 50% |
| 34 | Canada | English: 65%, French: 35% |
| 35 | Cape Verde | Portuguese: 100% |
| 36 | Cayman Islands | English: 95%, Spanish: 5% |
| 37 | Central African Republic | French: 100% |
| 38 | Chad | French: 50%, Arabic: 50% |
| 39 | Channel Islands | English: 100% |
| 40 | Chile | English: 100% |
| 41 | China | Chinese: 100% |
| 42 | Colombia | Spanish: 100% |
| 43 | Comoros | Arabic: 100% |
| 44 | Congo, Dem. Rep | French: 100% |
| 45 | Congo, Republic | French: 100% |
| 46 | Costa Rica | Spanish: 100% |
| 47 | Côte d'Ivoire | French: 100% |
| 48 | Croatia | Serbo-Croatian: 100% |
| 49 | Cuba | Spanish: 100% |
| 50 | Cyprus | Greek: 50%, French: 50% |
| 51 | Czech Republic | Czech: 100% |
| 52 | Denmark | Danish: 100% |
| 53 | Djibouti | French: 50%, Arabic: 50% |
| 54 | Dominica | English: 100% |
| 55 | Dominican Republic | Spanish: 100% |
| 56 | Ecuador | Spanish: 100% |
| 57 | Egypt | Arabic: 100% |
| 58 | El Salvador | Spanish: 100% |
| 59 | Equatorial Guinea | Spanish: 75%, French: 25% |
| 60 | Eritrea | Arabic: 70%, English: 30% |
| 61 | Estonia | Estonian: 70%, Russian: 30% |
| 62 | Ethiopia | Amharic: 32.7%, Oromo: 31.6%, Arabic: 7.5%, English: 7.5% |
| 63 | Faeroe Islands | Danish: 100% |
| 64 | Falkland Islands | English: 100% |
| 65 | Fiji | Hindi: 50%, English: 50% |
| 66 | Finland | Finnish: 95%, Swedish: 5% |
| 67 | France | French: 100% |
| 68 | French Guiana | French: 100% |
| 69 | French Polynesia | French: 61.1% |
| 70 | Gabon | French: 100% |
| 71 | Gambia | English: 100% |
| 72 | Georgia | Georgian: 80%, Russian: 10%, Armenian: 10% |
| 73 | Germany | German: 100% |
| 74 | Ghana | English: 100% |
| 75 | Gibraltar | English: 100% |
| 76 | Greece | Greek: 100% |
| 77 | Greenland | Danish: 13.7% |
| 78 | Grenada | English: 100% |
| 79 | Guam | English: 38.3%, Chamorro: 22.2%, Filipino: 22.2% |
| 80 | Guatemala | Spanish: 100% |
| 81 | Guinea | French: 100% |
| 82 | Guinea-Bissau | Portuguese: 100% |
| 83 | Guyana | English: 100% |
| 84 | Haiti | Haitian: 74.8%, French: 25.2% |
| 85 | Honduras | Spanish: 100% |
| 86 | Hong Kong | Chinese: 91.7%, English: 2.8% |
| 87 | Hungary | Hungarian: 93.6% |
| 88 | Iceland | Icelandic: 100% |
| 89 | India | Hindi: 41%, Bengali: 8.1%, Telugu: 7.2%, Marathi: 7%, Tamil: 5.9%, Urdu: 5%, Malayalam: 3.2%, Panjabi: 2.8% |
| 90 | Indonesia | Malay: 100% |
| 91 | Iran | Persian: 75%, Kurdish: 20%, Arabic: 5% |
| 92 | Iraq | Arabic: 80%, Kurdish: 20% |
| 93 | Ireland | English: 95%, Irish: 5% |
| 94 | Israel | Hebrew: 80%, Arabic: 20% |
| 95 | Italy | Italian: 100% |
| 96 | Jamaica | English: 100% |
| 97 | Japan | Japanese: 100% |
| 98 | Jordan | Arabic: 100% |
| 99 | Kazakhstan | Kazakh: 60%, Russian: 40% |
| 100 | Kenya | Swahili: 80%, English: 20% |
| 101 | Korea, DPR | Korean: 100% |
| 102 | Kuwait | Arabic: 50%, English: 50% |
| 103 | Kyrgyzstan | Kirghiz: 64.7%, Uzbek: 13.6%, Russian: 12.5% |
| 104 | Laos | Lao: 100% |
| 105 | Latvia | Latvian: 58.2%, Russian: 37.5%, Lithuanian: 4.3% |
| 106 | Lebanon | Arabic: 100% |
| 107 | Lesotho | English: 10% |
| 108 | Liberia | English: 20% |
| 109 | Libya | Arabic: 90%, English: 5%, Italian: 5% |
| 110 | Lithuania | Lithuanian: 80%, Russian: 10%, Polish: 10% |
| 111 | Luxembourg | Luxembourgish: 77%, French: 6%, German: 4%, English: 1% |
| 112 | Macao (China) | Chinese: 97%, English: 1.5%, Filipino: 1.3% |
| 113 | Macedonia | Macedonian: 66.5%, Albanian: 25.1%, Turkish: 8.4% |
| 114 | Madagascar | French: 100% |
| 115 | Malawi | English: 100% |
| 116 | Malaysia | Malay: 100% |
| 117 | Maldives | Dhivehi: 95%, English: 5% |
| 118 | Mali | Bambara: 80% |
| 119 | Malta | Maltese: 90.2%, English: 6% |
| 120 | Marshall Islands | English: 100% |
| 121 | Martinique | French: 100% |
| 122 | Mauritania | Arabic: 100% |
| 123 | Mauritius | Haitian: 80.5%, French: 3.5%, English: 1% |
| 124 | Mayotte | French: 100% |
| 125 | Mexico | Spanish: 100% |
| 126 | Micronesia | English: 100% |
| 127 | Moldova | Romanian: 75.17%, Russian: 15.99%, Ukrainian: 3.85%, Bulgarian: 1.14% |
| 128 | Mongolia | Mongolian: 100% |
| 129 | Montenegro | Serbo-Croatian: 91%, Albanian: 5% |
| 130 | Morocco | Arabic: 100% |
| 131 | Mozambique | Portuguese: 10.7% |
| 132 | Myanmar (Burma) | Burmese: 66.7% |
| 133 | Namibia | German: 32%, English: 7%, Afrikaans: 4.4% |
| 134 | Netherlands | Dutch: 100% |
| 135 | New Caledonia | French: 100% |
| 136 | New Zealand | English: 91.2% |
| 137 | Nicaragua | Spanish: 97.5% |
| 138 | Niger | French: 100% |
| 139 | Nigeria | English: 100% |
| 140 | Norfolk Island | English: 100% |
| 141 | Norway | Norwegian: 100% |
| 142 | Oman | Arabic: 100% |
| 143 | Pakistan | Panjabi: 48%, Urdu: 8% |
| 144 | Palau | Filipino: 13.5%, English: 9.4%, Chinese: 5.7%, Japanese: 1.5% |
| 145 | Palestinian State | Arabic: 100% |
| 146 | Panama | Spanish: 86%, English: 14% |
| 147 | Papua New Guinea | English: 1% |
| 148 | Paraguay | Spanish: 3.1% |
| 149 | Peru | Spanish: 84.1% |
| 150 | Philippines | Filipino: 55%, English: 4% |
| 151 | Poland | Polish: 97.8% |
| 152 | Portugal | Portuguese: 100% |
| 153 | Puerto Rico | Spanish: 87%, English: 2.5% |
| 154 | Qatar | Arabic: 50%, English: 50% |
| 155 | Korea, Republic | Korean: 100% |
| 156 | Reunion | French: 100% |
| 157 | Romania | Romanian: 91%, Hungarian: 6.7% |
| 158 | Russia | Russian: 100% |
| 159 | Rwanda | French: 20%, English: 20% |
| 160 | Saint Pierre and Miquelon | French: 100% |
| 161 | Sao Tome and Principe | Portuguese: 100% |
| 162 | Saudi Arabia | Arabic: 100% |
| 163 | Senegal | French: 100% |
| 164 | Serbia | Serbo-Croatian: 95%, Hungarian: 5% |
| 165 | Seychelles | Haitian: 95%, English: 5% |
| 166 | Sierra Leone | English: 100% |
| 167 | Singapore | Chinese: 40.7%, English: 23%, Malay: 14.1% |
| 168 | Slovakia | Slovak: 90%, Hungarian: 10% |
| 169 | Slovenia | Slovenian: 100% |
| 170 | Solomon Islands | English: 2% |
| 171 | Somalia | Somali: 80%, Arabic: 10%, English: 5%, Italian: 5% |
| 172 | South Africa | Zulu: 23.83%, Xhosa: 17.64%, Afrikaans: 13.35%, Pedi: 9.39%, English: 8.2% |
| 173 | South Sudan | Arabic: 50%, English: 50% |
| 174 | Spain | Spanish: 74%, Catalan: 17%, Galician: 7%, Basque: 2% |
| 175 | Sri Lanka | Sinhala: 74%, Tamil: 18% |
| 176 | St. Helena | English: 100% |
| 177 | St. Kitts and Nevis | English: 100% |
| 178 | St. Lucia | English: 100% |
| 179 | St. Vincent and the Grenadines | English: 100% |
| 180 | Sudan | Arabic: 50%, English: 50% |
| 181 | Suriname | Dutch: 100% |
| 182 | Swaziland | English: 100% |
| 183 | Sweden | Swedish: 100% |
| 184 | Switzerland | German: 66.7%, French: 23.4%, Italian: 8.9%, English: 1% |
| 185 | Syria | Arabic: 100% |
| 186 | Taiwan | Chinese: 100% |
| 187 | Tajikistan | Tajik: 100% |
| 188 | Tanzania | Swahili: 90%, English: 10% |
| 189 | Thailand | Thai: 100% |
| 190 | Togo | French: 100% |
| 191 | Trinidad and Tobago | English: 100% |
| 192 | Tunisia | Arabic: 100% |
| 193 | Turkey | Turkish: 90%, Kurdish: 6%, Arabic: 1.2% |
| 194 | Turkmenistan | Turkmen: 72%, Russian: 12%, Uzbek: 9% |
| 195 | Turks and Caicos Islands | English: 100% |
| 196 | Uganda | Swahili: 80%, English: 20% |
| 197 | Ukraine | Ukrainian: 67%, Russian: 24% |
| 198 | United Arab Emirates | Arabic: 100% |
| 199 | United Kingdom | English: 96.3%, Scottish Gaelic: 2.5%, Welsh: 1.2% |
| 200 | United States | English: 82.1%, Spanish: 10.7% |
| 201 | Uruguay | Spanish: 100% |
| 202 | Uzbekistan | Uzbek: 74.3%, Russian: 14.2%, Tajik: 4.4% |
| 203 | Vanuatu | English: 2% |
| 204 | Venezuela | Spanish: 100% |
| 205 | Vietnam | Vietnamese: 100% |
| 206 | Virgin Islands, U.S. | English: 74.7%, Spanish: 16.8%, French: 6.6% |
| 207 | Wallis and Futuna | French: 10.8% |
| 208 | Western Sahara | Arabic: 100% |
| 209 | Yemen | Arabic: 100% |
| 210 | Zambia | English: 1.7% |
| 211 | Zimbabwe | English: 100% |

Table 4.2: Language demographics by country. Values for each country add to 100% or less.

was born in Thessaloniki, present-day Greece, and would earn points for Greek. Because our language distribution statistics are from the last few years, we include only people born in 1800 and later, to reduce the effect of geopolitical and cultural changes on our mapping of countries to languages. To match the year limitation of the Human Accomplishment dataset, I also set 1950 as the latest year of birth for the Wikipedia dataset.

Despite some inaccuracies, using present-day countries provides a consistent mapping people who lived over a period of several millennia to their contemporary countries. Moreover, using present-day countries allows me to use the present-day language distribution statistics for each country to identify the main languages spoken in a country and determine the language affiliation of each person.

### 4.1.2 Wikipedia

Wikipedia is available in more than 270 language editions. As Wikipedia is collaboratively authored, each edition reflects the knowledge of the language community that contributed to it [27, 32]. For example, an article about Plato in the Filipino Wikipedia indicates that Plato is known enough among speakers of Filipino to motivate some of them to write an article about him. While a Wikipedia article in just one language can be the result of short-lived fame within a limited community, a person with articles written about him or her in many languages has likely made a substantial cultural contribution that impacted people from a diverse linguistic and cultural background.

I compiled the Wikipedia dataset of illustrious people as follows. I started by retrieving a table of 2,345,208 people from Freebase (`www.freebase.com`), a collaboratively curated repository of structured data of millions of entities, such places and people. I used a data dump from November 4, 2012; the latest version of the table is available at [22]. For each person, the table contains his or her name, date of birth, place of birth, occupation, and additional information. In addition, for each person with an article in the English Wikipedia, Freebase stores the Wikipedia unique identifier (known as *pageid* or *curid*) of the respective article, which I retrieved through the Freebase API [23]. The pageid and the Wikipedia API [75] were used to find the number of language editions in which a person

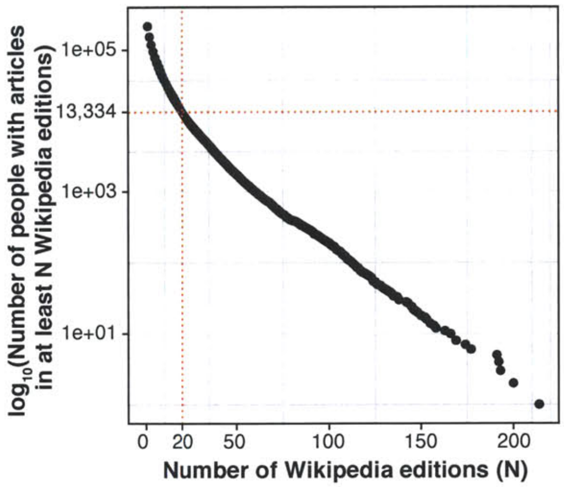


Figure 4.1: Number of biographical articles with versions in at least $N$ Wikipedia language editions.

had an article. Then, the pageid, Wikipedia article name, and number of languages of each article were added to the table retrieved from Freebase.

I matched 991,684 people with the English Wikipedia, from which I selected 216,280 people with a defined date of birth, place of birth and gender. I then restricted this list to include only the 13,334 people who had articles in at least 20 Wikipedia language editions. The 20-language threshold generated a group that is exclusive enough while still containing enough data points (Figure 4.1). I refer to this dataset as *Wikipedia 20*. For comparison, a 25-language threshold would give 8,942 articles, and a 30-language threshold only 6,336.

Next, I converted dates to a standard four-digit year format. While doing so, I fixed all BCE years, which the Freebase dump listed one year off. For example, Jesus's year of birth was listed as 3 BCE instead of 4 BCE. I then used the Google Geocoding API [26] to resolve the listed places of birth to latitude-longitude coordinates, and used the GeoNames database (`www.geonames.com`) to resolve the coordinates into the present-day name of the country in which each person was born. Finally, I converted countries to languages as described in Section 4.1.1 above. To increase the accuracy of the conversion, I selected from the Wikipedia 20 dataset only the 6,158 people who were born after 1800 and before 1950. Tables 4.3 and 4.4 show the number of illustrious people for each country and language, respectively.

| | Country | People (all years) | People (1800-1950) |
|---|---|---|---|
| 1 | Afghanistan | 13 | 6 |
| 2 | Albania | 20 | 11 |
| 3 | Algeria | 21 | 11 |
| 4 | Andorra | 2 | 1 |
| 5 | Angola | 6 | 4 |
| 6 | Argentina | 124 | 39 |
| 7 | Armenia | 14 | 7 |
| 8 | Aruba | 1 | 1 |
| 9 | Australia | 153 | 42 |
| 10 | Austria | 165 | 111 |
| 11 | Azerbaijan | 14 | 11 |
| 12 | Bahrain | 2 | 1 |
| 13 | Bangladesh | 9 | 8 |
| 14 | Barbados | 2 | N/A |
| 15 | Belarus | 27 | 11 |
| 16 | Belgium | 109 | 49 |
| 17 | Belize | 2 | 1 |
| 18 | Benin | 5 | 1 |
| 19 | Bermuda | 1 | N/A |
| 20 | Bolivia | 4 | 2 |
| 21 | Bosnia and Herzegovina | 33 | 9 |
| 22 | Botswana | 4 | 3 |
| 23 | Brazil | 165 | 59 |
| 24 | Brunei | 1 | 1 |
| 25 | Bulgaria | 24 | 11 |
| 26 | Burkina Faso | 3 | 2 |
| 27 | Burundi | 2 | N/A |
| 28 | Cambodia | 6 | 5 |
| 29 | Cameroon | 16 | 1 |
| 30 | Canada | 192 | 77 |
| 31 | Cape Verde | 4 | 2 |
| 32 | Central African Republic | 6 | 4 |
| 33 | Chad | 3 | 1 |
| 34 | Chile | 36 | 18 |
| 35 | China | 120 | 46 |
| 36 | Colombia | 24 | 5 |
| 37 | Comoros | 1 | N/A |
| 38 | Congo, Republic | 8 | 1 |
| 39 | Costa Rica | 5 | 3 |
| 40 | Côte d'Ivoire | 8 | 1 |
| 41 | Croatia | 58 | 11 |
| 42 | Cuba | 19 | 14 |
| 43 | Cyprus | 10 | 7 |
| 44 | Czech Republic | 142 | 70 |
| 45 | Congo, Dem. Rep. | 6 | 3 |
| 46 | Denmark | 96 | 38 |
| 47 | Djibouti | 2 | 1 |
| 48 | Dominica | 2 | 1 |
| 49 | Dominican Republic | 4 | 1 |
| 50 | East Timor | 3 | 3 |
| 51 | Ecuador | 6 | N/A |
| 52 | Egypt | 37 | 20 |
| 53 | El Salvador | 3 | 1 |
| 54 | Equatorial Guinea | 1 | 1 |
| 55 | Eritrea | 3 | 1 |
| 56 | Estonia | 23 | 11 |
| 57 | Ethiopia | 7 | 2 |
| 58 | Faroe Islands | 4 | 2 |
| 59 | Fiji | 2 | 1 |
| 60 | Finland | 84 | 39 |
| 61 | France | 930 | 491 |
| 62 | French Guiana | 2 | 1 |
| 63 | French Polynesia | 1 | N/A |
| 64 | Gabon | 3 | 3 |
| 65 | Georgia | 32 | 14 |
| 66 | Germany | 798 | 437 |
| 67 | Ghana | 25 | 5 |
| 68 | Gibraltar | 1 | N/A |
| 69 | Greece | 105 | 33 |
| 70 | Greenland | 3 | 1 |
| 71 | Grenada | 2 | 1 |
| 72 | Guadeloupe | 4 | 1 |
| 73 | Guam | 1 | 1 |
| 74 | Guatemala | 5 | 3 |
| 75 | Guernsey | 1 | N/A |
| 76 | Guinea | 3 | 2 |
| 77 | Guinea-Bissau | 6 | 5 |
| 78 | Guyana | 1 | N/A |
| 79 | Haiti | 7 | 1 |
| 80 | Honduras | 7 | 1 |
| 81 | Hong Kong | 14 | 3 |
| 82 | Hungary | 86 | 67 |
| 83 | Iceland | 14 | 5 |
| 84 | India | 165 | 88 |
| 85 | Indonesia | 11 | 9 |
| 86 | Iran | 47 | 17 |
| 87 | Iraq | 19 | 6 |
| 88 | Ireland | 110 | 40 |
| 89 | Isle of Man | 4 | 3 |
| 90 | Israel | 56 | 28 |
| 91 | Italy | 644 | 220 |
| 92 | Jamaica | 18 | 7 |
| 93 | Japan | 216 | 93 |
| 94 | Jersey | 2 | N/A |
| 95 | Jordan | 5 | 2 |
| 96 | Kazakhstan | 27 | 20 |
| 97 | Kenya | 19 | 8 |
| 98 | Kosovo | 4 | N/A |
| 99 | Kuwait | 3 | 3 |
| 100 | Kyrgyzstan | 3 | 2 |
| 101 | Latvia | 17 | 12 |
| 102 | Lebanon | 21 | 9 |
| 103 | Lesotho | 1 | 1 |
| 104 | Liberia | 5 | 3 |
| 105 | Libya | 6 | 1 |
| 106 | Lithuania | 25 | 15 |
| 107 | Luxembourg | 10 | 5 |
| 108 | Macedonia | 15 | 3 |
| 109 | Madagascar | 4 | 3 |
| 110 | Malawi | 4 | 3 |
| 111 | Malaysia | 10 | 4 |
| 112 | Maldives | 3 | 1 |
| 113 | Mali | 11 | 7 |
| 114 | Malta | 7 | 4 |
| 115 | Martinique | 2 | 2 |
| 116 | Mauritania | 3 | 3 |
| 117 | Mauritius | 2 | 2 |
| 118 | Mexico | 68 | 25 |
| 119 | Moldova | 7 | 1 |
| 120 | Monaco | 2 | 1 |
| 121 | Mongolia | 7 | 2 |
| 122 | Montenegro | 9 | 2 |
| 123 | Morocco | 20 | 10 |
| 124 | Mozambique | 7 | 2 |
| 125 | Myanmar (Burma) | 9 | 9 |
| 126 | Namibia | 4 | 3 |
| 127 | Nauru | 3 | 2 |
| 128 | Nepal | 2 | N/A |
| 129 | Netherlands | 205 | 76 |
| 130 | New Caledonia | 1 | N/A |
| 131 | New Zealand | 25 | 10 |
| 132 | Nicaragua | 7 | 7 |
| 133 | Niger | 5 | 3 |
| 134 | Nigeria | 37 | 6 |
| 135 | Korea, DPR | 7 | 4 |
| 136 | Norway | 79 | 42 |
| 137 | Oman | 2 | N/A |
| 138 | Pakistan | 31 | 18 |
| 139 | Palau | 2 | 1 |
| 140 | Palestinian State | 3 | 1 |
| 141 | Panama | 5 | 3 |
| 142 | Papua New Guinea | 1 | 1 |
| 143 | Paraguay | 11 | 2 |
| 144 | Peru | 22 | 15 |
| 145 | Philippines | 25 | 19 |
| 146 | Poland | 183 | 125 |
| 147 | Portugal | 84 | 21 |
| 148 | Puerto Rico | 12 | 3 |
| 149 | Qatar | 1 | N/A |
| 150 | Romania | 65 | 34 |
| 151 | Russia | 393 | 249 |
| 152 | Rwanda | 2 | 1 |
| 153 | Saint Kitts and Nevis | 3 | N/A |
| 154 | Saint Lucia | 1 | 1 |
| 155 | Saint Vincent and the Grenadines | 1 | 1 |
| 156 | Samoa | 2 | 2 |
| 157 | Saudi Arabia | 14 | 6 |
| 158 | Senegal | 12 | 3 |
| 159 | Serbia | 68 | 17 |
| 160 | Seychelles | 2 | 2 |
| 161 | Sierra Leone | 3 | 1 |
| 162 | Singapore | 7 | 4 |
| 163 | Slovakia | 34 | 10 |
| 164 | Slovenia | 28 | 6 |
| 165 | Solomon Islands | 1 | N/A |
| 166 | Somalia | 11 | 5 |
| 167 | South Africa | 64 | 26 |
| 168 | Korea, Republic | 42 | 10 |
| 169 | South Sudan | 1 | 1 |
| 170 | Spain | 346 | 100 |
| 171 | Sri Lanka | 7 | 6 |
| 172 | St. Lucia | 1 | 1 |
| 173 | Sudan | 4 | 1 |
| 174 | Suriname | 7 | 3 |
| 175 | Sweden | 175 | 72 |
| 176 | Switzerland | 111 | 57 |
| 177 | Syria | 10 | 1 |
| 178 | Taiwan | 15 | 4 |
| 179 | Tajikistan | 4 | 1 |
| 180 | Tanzania | 4 | 3 |
| 181 | Thailand | 6 | 4 |
| 182 | Bahamas | 4 | 2 |
| 183 | Gambia | 1 | N/A |
| 184 | Togo | 6 | 3 |
| 185 | Tonga | 2 | 1 |
| 186 | Trinidad and Tobago | 6 | 2 |
| 187 | Tunisia | 11 | 4 |
| 188 | Turkey | 96 | 27 |
| 189 | Turkmenistan | 3 | 1 |
| 190 | Virgin Islands, U.S. | 2 | 1 |
| 191 | Uganda | 3 | 2 |
| 192 | Ukraine | 114 | 62 |
| 193 | United Arab Emirates | 4 | 4 |
| 194 | United Kingdom | 1515 | 673 |
| 195 | United States | 3726 | 1807 |
| 196 | Uruguay | 35 | 8 |
| 197 | Uzbekistan | 9 | 2 |
| 198 | Venezuela | 13 | 2 |
| 199 | Vietnam | 12 | 10 |
| 200 | Yemen | 1 | 1 |
| 201 | Zambia | 2 | 2 |
| 202 | Zimbabwe | 7 | 4 |
| | *Total* | *13334* | *6158* |

Table 4.3: Number of people with articles in at least 20 Wikipedia language editions, by country.

| | Language | Code | People (all years) | People (1800-1950) | | Language | Code | People (all years) | People (1800-1950) |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Afrikaans | **afr** | **8.7** | **3.6** | 33 | Korean | **kor** | **7** | **4** |
| 2 | Albanian | **sqi** | **21.5** | **11.1** | 34 | Latvian | **lav** | **9.9** | **7** |
| 3 | Arabic | **ara** | **194.2** | **86.9** | 35 | Lithuanian | **lit** | **20.7** | **12.5** |
| 4 | Armenian | **hye** | **17.1** | **8.4** | 36 | Macedonian | **mkd** | **5.3** | **1.3** |
| 5 | Azerbaijani | **aze** | **12.6** | **9.9** | 37 | Malay | **msa** | **22.8** | **14.4** |
| 6 | Basque | **eus** | **6.9** | **2** | 38 | Malayalam | **mal** | **5.3** | **2.8** |
| 7 | Belarusian | **bel** | **9.9** | **4** | 39 | Maltese | **mlt** | **6.3** | **3.6** |
| 8 | Bengali | **ben** | **22.2** | **15** | 40 | Marathi | **mar** | **11.6** | **6.2** |
| 9 | Bulgarian | **bul** | **20.4** | **9.3** | 41 | Mongolian | **mon** | **7** | **2** |
| 10 | Catalan | **cat** | **59.8** | **17.5** | 42 | Norwegian | **nor** | **79** | **42** |
| 11 | Chinese | **zho** | **154.6** | **55.5** | 43 | Persian | **fas** | **41.8** | **15.8** |
| 12 | Czech | **ces** | **37** | **16** | 44 | Polish | **pol** | **181.5** | **123.8** |
| 13 | Danish | **dan** | **96.4** | **38.1** | 45 | Portuguese | **por** | **264.6** | **90.4** |
| 14 | Dutch | **nld** | **267.5** | **106.5** | 46 | Romanian | **ron** | **65.4** | **32.2** |
| 15 | English | **eng** | **5071** | **2325.1** | 47 | Russian | **rus** | **449.6** | **279.6** |
| 16 | Estonian | **est** | **16.1** | **7.7** | 48 | Serbo-Croatian | **hbs** | **145.4** | **34.4** |
| 17 | Filipino | **fil** | **14.2** | **10.8** | 49 | Sinhala | **sin** | **5.2** | **4.4** |
| 18 | Finnish | **fin** | **79.8** | **37** | 50 | Slovak | **slk** | **30.6** | **9** |
| 19 | French | **fra** | **1145.8** | **586.1** | 51 | Slovenian | **slv** | **28** | **6** |
| 20 | Galician | **glg** | **24.2** | **7** | 52 | Spanish | **spa** | **994.6** | **391.1** |
| 21 | Georgian | **kat** | **25.6** | **11.2** | 53 | Swahili | **swa** | **21.2** | **10.7** |
| 22 | German | **deu** | **1019.9** | **574.5** | 54 | Swedish | **swe** | **179.2** | **74** |
| 23 | Greek | **ell** | **112** | **37** | 55 | Tajik | **tgk** | **4.4** | **1.1** |
| 24 | Haitian | **hat** | **9.4** | **4.6** | 56 | Tamil | **tam** | **11** | **6.3** |
| 25 | Hebrew | **heb** | **44.8** | **22.4** | 57 | Thai | **tha** | **6** | **4** |
| 26 | Hindi | **hin** | **68.7** | **36.6** | 58 | Turkish | **tur** | **93.2** | **28.1** |
| 27 | Hungarian | **hun** | **91.7** | **66.8** | 59 | Turkmen | **tuk** | **2.2** | **0.7** |
| 28 | Icelandic | **isl** | **14** | **5** | 60 | Ukrainian | **ukr** | **76.6** | **41.6** |
| 29 | Italian | **ita** | **661.9** | **227.5** | 61 | Urdu | **urd** | **10.7** | **5.8** |
| 30 | Japanese | **jpn** | **216** | **93** | 62 | Uzbek | **uzb** | **7.4** | **1.8** |
| 31 | Kazakh | **kaz** | **16.2** | **12** | 63 | Vietnamese | **vie** | **12** | **10** |
| 32 | Kirghiz | **kir** | **1.9** | **1.3** | 64 | Welsh | **cym** | **18.2** | **8.1** |

Table 4.4: Number of people with articles in at least 20 Wikipedia language editions, by language (rounded to the nearest tenth).

### 4.1.3 *Human Accomplishment*

The second measure of illustrious people is based on the book *Human Accomplishment: The Pursuit of Excellence in the Arts and Sciences, 800 B.C. to 1950* [45], which ranks the contribution of 3,869 people to different fields of arts and science. Each listed person is ranked on a scale of 1 to 100 for his or her contribution to one or more of the following fields: art, literature, music, philosophy, astronomy, biology, chemistry, earth sciences, mathematics, medicine, physics and technology. People who contributed to more than one field were ranked separately for each field. For example, Isaac Newton received the highest score of 100 for his contribution in physics, and a score of 88.93 for his contribution in mathematics. For each person, the Human Accomplishment tables contain his or her name, ranking in all relevant fields, year of birth, year of death, year flourished, country of birth and country of work. I considered each person that was listed on Human Accomplishment an illustrious person, regardless of his or her rank.

To find the number of notable people for each language group, I converted countries of birth to languages as explained in Section 4.1.1. In most cases, I used the countries of birth as listed on Human Accomplishment. However, the dataset occasionally provided a geographical or cultural region, rather than a country, as a place of birth: *Balkans, Latin America, Sub-Saharan Africa, Arab World, Ancient Greece* and *Rome*. I replaced the first three with the specific places of birth for the respective people, as listed on Wikipedia 20, and converted them to languages based on their present-day countries. I did not resolve *Arab World, Ancient Greece* or *Rome* to specific locations, but instead converted them directly to *Arabic, Ancient Greek*, or *Latin*, respectively. As with the Wikipedia 20 dataset, I increased the accuracy of the country-to-language mapping by selecting only the 1,655 people born between 1800 and 1950. Tables 4.5 and 4.6 show the number of illustrious people for each country and language, respectively.

## 4.2 Results

Figure 4.2 shows the bivariate correlation between the number of illustrious people measured using the Wikipedia dataset and the eigenvector centrality of that language in the

| | Country | People (all years) | People (1800-1950) |
|---|---|---|---|
| 1 | *Ancient Greece* | 134 | N/A |
| 2 | *Arab World* | 86 | 14 |
| 3 | Argentina | 2 | 2 |
| 4 | Australia | 4 | 4 |
| 5 | Austria | 75 | 48 |
| 6 | Belgium | 82 | 27 |
| 7 | Brazil | 3 | 3 |
| 8 | Bulgaria | 1 | 1 |
| 9 | Canada | 11 | 11 |
| 10 | Chile | 3 | 3 |
| 11 | China | 237 | 22 |
| 12 | Croatia | 5 | 3 |
| 13 | Cuba | 3 | 3 |
| 14 | Czech Republic | 48 | 28 |
| 15 | Denmark | 37 | 20 |
| 16 | Finland | 6 | 5 |
| 17 | France | 542 | 236 |
| 18 | Germany | 536 | 267 |
| 19 | Greece | 9 | 6 |
| 20 | Guatemala | 1 | 1 |
| 21 | Hungary | 21 | 18 |
| 22 | Iceland | 2 | 1 |
| 23 | India | 93 | 16 |
| 24 | Italy | 389 | 58 |
| 25 | Japan | 169 | 57 |
| 26 | Kenya | 1 | 1 |
| 27 | Mexico | 5 | 4 |
| 28 | Montenegro | 1 | 1 |
| 29 | Netherlands | 84 | 31 |
| 30 | New Zealand | 3 | 3 |
| 31 | Nicaragua | 1 | 1 |
| 32 | Norway | 23 | 22 |
| 33 | Peru | 1 | 1 |
| 34 | Poland | 25 | 21 |
| 35 | Portugal | 11 | 4 |
| 36 | Romania | 5 | 4 |
| 37 | *Rome* | 55 | N/A |
| 38 | Russia | 134 | 118 |
| 39 | Serbia | 2 | 2 |
| 40 | Slovakia | 4 | 4 |
| 41 | Slovenia | 2 | 2 |
| 42 | South Africa | 1 | 1 |
| 43 | Spain | 76 | 26 |
| 44 | Sweden | 44 | 21 |
| 45 | Switzerland | 64 | 32 |
| 46 | United Kingdom | 531 | 230 |
| 47 | United States | 297 | 272 |
| | *Total* | *3869* | *1655* |

Table 4.5: Number of people listed on *Human Accomplishment*, by country.

| | Language | Code | People (all years) | People (1800-1950) |
|---|---|---|---|---|
| 1 | Afrikaans | afr | 0.1 | 0.1 |
| 2 | Albanian | sqi | 0 | 0 |
| 3 | Arabic | ara | 86 | 14 |
| 4 | Basque | eus | 1.5 | 0.5 |
| 5 | Bengali | ben | 7.5 | 1.3 |
| 6 | Bulgarian | bul | 0.8 | 0.8 |
| 7 | Catalan | cat | 12.9 | 4.4 |
| 8 | Chinese | zho | 237.1 | 22.1 |
| 9 | Czech | ces | 48 | 28 |
| 10 | Danish | dan | 37 | 20 |
| 11 | Dutch | nld | 133.2 | 47.2 |
| 12 | English | eng | 772.1 | 461.4 |
| 13 | Finnish | fin | 5.7 | 4.8 |
| 14 | French | fra | 593.6 | 258.1 |
| 15 | Galician | glg | 5.3 | 1.8 |
| 16 | German | deu | 645.1 | 330.9 |
| 17 | Greek | ell | 8.1 | 5.1 |
| 18 | Hindi | hin | 38.1 | 6.6 |
| 19 | Hungarian | hun | 20.5 | 17.6 |
| 20 | Icelandic | isl | 2 | 1 |
| 21 | Italian | ita | 394.8 | 61 |
| 22 | Japanese | jpn | 169 | 57 |
| 23 | Malayalam | mal | 3 | 0.5 |
| 24 | Marathi | mar | 6.5 | 1.1 |
| 25 | Norwegian | nor | 23 | 22 |
| 26 | Polish | pol | 24.4 | 20.5 |
| 27 | Portuguese | por | 14 | 7 |
| 28 | Romanian | ron | 4.5 | 3.6 |
| 29 | Russian | rus | 134 | 118 |
| 30 | Serbo-Croatian | hbs | 9.5 | 6.9 |
| 31 | Slovak | slk | 3.6 | 3.6 |
| 32 | Slovenian | slv | 2 | 2 |
| 33 | Spanish | spa | 100.5 | 59.9 |
| 34 | Swahili | swa | 0.8 | 0.8 |
| 35 | Swedish | swe | 44.3 | 21.2 |
| 36 | Tamil | tam | 5.5 | 0.9 |
| 37 | Turkish | tur | 1.8 | 1.2 |
| 38 | Urdu | urd | 4.7 | 0.8 |
| 39 | Welsh | cym | 6.4 | 2.8 |

Table 4.6: Number of people listed on *Human Accomplishment*, by language (rounded to the nearest tenth).

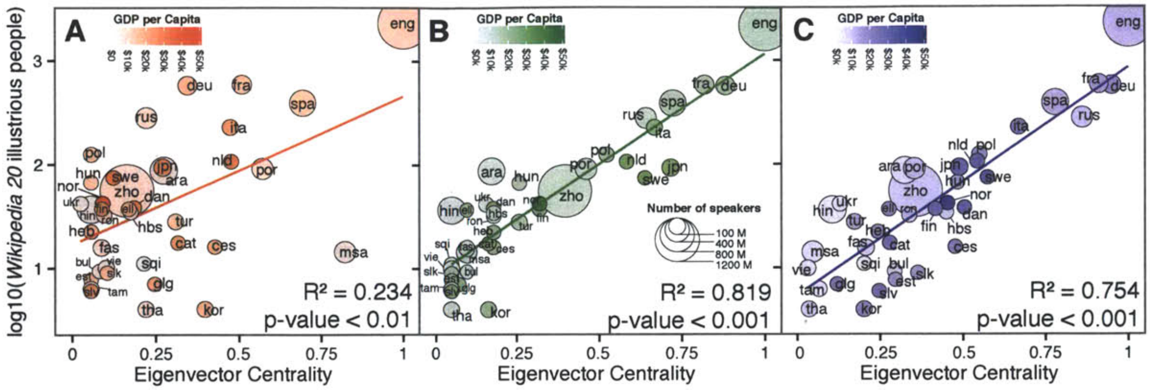

Figure 4.2: Number of people per language (born 1800-1950) with articles in at least 20 Wikipedia language editions as a function of their language's eigenvector centrality in the **A** Twitter GLN, **B** Wikipedia GLN, and **C** book translations GLN. Circle size represents the number of speakers for each language, and the color intensity represents GDP per capita for the language.

Twitter, Wikipedia and book translation networks. I use only the 38 languages that are present in all three GLNs. Table 4.7 presents these results in the form of a regression table where variables are introduced sequentially.

With the exception of the Twitter dataset, the correlation between the number of illustrious people and the eigenvector centrality of a language is higher than the correlation observed between the number of illustrious people and the income and population of the language group. In fact, although there is an important collinear component between the centrality of a language in the Wikipedia or book translation network and the income and population of its speakers, the orthogonal component explains an important amount of the variance. The semi-partial correlation, defined as the difference between the $R^2$ obtained from a regression with all variables and a regression where the variable in question has been removed, indicates that the percentage of the variance in the number of illustrious people explained by the Wikipedia and book translation GLNs are respectively 33.6% (F=77.57, p-value<0.001) and 35.5% (F=93.79, p-value<0.001) after the effects of income and population have been taken into account. In contrast, the semi-partial contribution of income and population is only 2.4% (F=2.82, p-value=0.07) when measured against the Wikipedia GLN, and 10.6% (F=14.1, p-val.ue<0.001) when measured against the book translation GLN. Results for all years are presented in Appendix C.

**A**

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| | Number of illustrious people born 1800-1950 per language, based on having biographies in over 20 Wikiepdia language editions | | | | | | |
| $\log_{10}$(Population) | 0.639*** | | | | 0.579*** | 0.033 | 0.294*** |
| | (0.109) | | | | (0.140) | (0.093) | (0.068) |
| $\log_{10}$(GDP per capita) | 0.996*** | | | | 0.922** | -0.261 | 0.059 |
| | (0.251) | | | | (0.274) | (0.203) | (0.165) |
| EV centrality [Twitter] | | 1.429** | | | 0.297 | | |
| | | (0.407) | | | (0.431) | | |
| EV centrality [Wikipedia] | | | 2.125*** | | | 2.196*** | |
| | | | (0.164) | | | (0.253) | |
| EV centrality [book trans.] | | | | 2.190*** | | | 1.928*** |
| | | | | (0.205) | | | (0.202) |
| (Intercept) | -3.559** | 1.224*** | 0.940*** | 0.740*** | -3.233* | 1.941* | 0.121 |
| | (1.126) | (0.139) | (0.067) | (0.095) | (1.229) | (0.898) | (0.709) |
| Observations | 38 | 38 | 38 | 38 | 38 | 38 | 38 |
| p-value | 0 | 0.001 | 0 | 0 | 0 | 0 | 0 |
| R-squared | 0.512 | 0.255 | 0.824 | 0.761 | 0.519 | 0.848 | 0.867 |
| Adjusted R-squared | 0.484 | 0.234 | 0.819 | 0.754 | 0.476 | 0.835 | 0.856 |

***,**,* significant at 0.1%, 1% and 5% levels, respectively. Standard errors in parentheses.

**B**

| Illustrious people by country | |
|---|---|
| United States | 1807 |
| United Kingdom | 673 |
| France | 491 |
| Germany | 437 |
| Russia | 238 |
| Italy | 326 |
| Poland | 125 |
| Austria | 111 |
| Spain | 100 |
| Japan | 93 |

**C**

| Illustrious people by language | |
|---|---|
| English | 2325.1 |
| French | 586.1 |
| German | 574.5 |
| Spanish | 391.1 |
| Russian | 279.6 |
| Italian | 227.5 |
| Polish | 123.8 |
| Dutch | 106.5 |
| Japanese | 93.0 |
| Portuguese | 90.4 |

**D**

| Twitter EV Cent. | |
|---|---|
| English | 1.00 |
| Malay | 0.82 |
| Spanish | 0.69 |
| Portuguese | 0.57 |
| French | 0.51 |
| Dutch | 0.48 |
| Italian | 0.47 |

**E**

| Wikipedia EV Cent. | |
|---|---|
| English | 1.00 |
| German | 0.88 |
| French | 0.82 |
| Spanish | 0.72 |
| Japanese | 0.72 |
| Italian | 0.67 |
| Russian | 0.64 |

**F**

| Book translation EV Cent. | |
|---|---|
| English | 1.00 |
| German | 0.95 |
| French | 0.91 |
| Russian | 0.86 |
| Spanish | 0.78 |
| Italian | 0.67 |
| Swedish | 0.57 |

Table 4.7: GLN centrality and the number of illustrious people per language according to *Wikipedia 20*. **A** Regression table explaining the number of people (born 1800-1950) of each language group about which there are articles in at least 20 Wikipedia language editions as a function of the language group's GDP per capita, population, and eigenvector (EV) centrality in each of GLNs. Cultural production rankings: the **B** countries and **C** languages that produced the largest number of people about which there are articles in at least 20 Wikipedia editions. GLN eigenvector centrality rankings for languages represented in biographies list: top seven languages in **D** the Twitter GLN, **E** the Wikipedia GLN, and **F** the book translation GLN.

Figure 4.3 and Table 4.8 show the same analysis but using the list of illustrious people from Human Accomplishment. I used only the languages that are present in all three GLNs. In addition, I removed Albanian as it proved to be a major outlier and the number of illustrious people associated with this language was negligible (0.05). The cultural influence of the languages as reflected in this biographical dataset is best explained by a combination of population, GDP and the centrality of a language in the book translation network (Table 4.8), which accounts for 91% of the variance. Centrality in the Wikipedia GLN or book translation GLN alone explains 76% and 84% of the variance, respectively, and 11.2% (F=13.07, p-value<0.001) and 24.9% (F=73.84, p-value<0.001) at the margin, as measured by the semi-partial correlation. Results for all years and results with Albanian are presented in Appendix C.

The data cannot distinguish between the hypothesis that speakers translate material from a hub language into their own language because the content produced in the hub language is more noteworthy, and the hypothesis that a person has an advantage in the competition for international prominence if he or she is born in a location associated with a

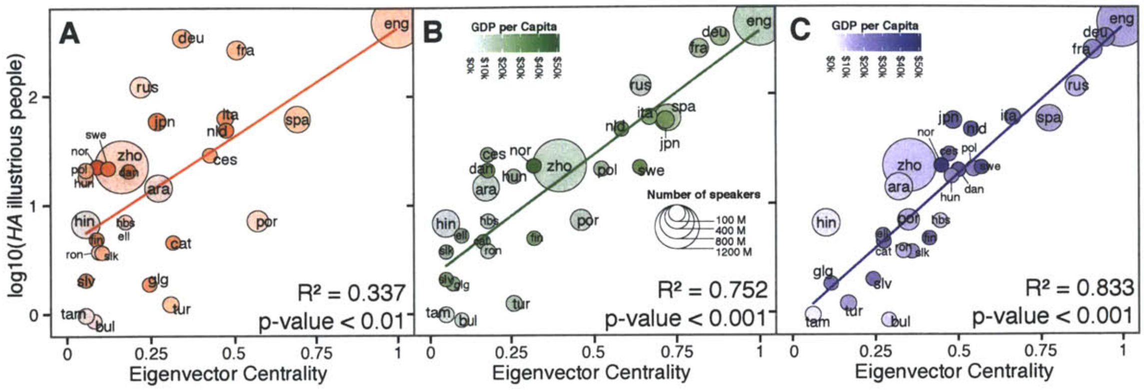


Figure 4.3: Number of people per language (born 1800-1950) listed in *Human Accomplishment* as a function of their language's eigenvector centrality in the **A** Twitter GLN, **B** Wikipedia GLN, and **C** book translations GLN. Circle size represents the number of speakers for each language, and the color intensity represents GDP per capita for the language.

**A**

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| | Number of illustrious people born 1800-1950 per language, based on inclusion in *Human Accomplishment* | | | | | | |
| $\log_{10}$(Population) | 0.874*** | | | | 0.800*** | 0.262 | 0.398*** |
| | (0.127) | | | | (0.176) | (0.202) | (0.087) |
| $\log_{10}$(GDP per capita) | 1.898*** | | | | 1.767*** | 0.568 | 0.768** |
| | (0.340) | | | | (0.404) | (0.470) | (0.222) |
| EV centrality [Twitter] | | 1.983*** | | | 0.322 | | |
| | | (0.508) | | | (0.521) | | |
| EV centrality [Wikipedia] | | | 2.253*** | | | 1.710** | |
| | | | (0.243) | | | (0.482) | |
| EV centrality [book trans.] | | | | 2.720*** | | | 2.006*** |
| | | | | (0.229) | | | (0.238) |
| (Intercept) | -8.262*** | 0.622** | 0.316** | -0.103 | -7.678*** | -2.299 | -3.657*** |
| | (1.561) | (0.176) | (0.113) | (0.120) | (1.841) | (2.125) | (0.979) |
| Observations | 29 | 29 | 29 | 29 | 29 | 29 | 29 |
| p-value | 0 | 0.001 | 0 | 0 | 0 | 0 | 0 |
| R-squared | 0.664 | 0.361 | 0.761 | 0.839 | 0.669 | 0.776 | 0.913 |
| Adjusted R-squared | 0.638 | 0.337 | 0.752 | 0.833 | 0.629 | 0.75 | 0.902 |

***,**,* significant at 0.1%, 1% and 5% levels, respectively. Standard errors in parentheses.

**B** Illustrious people by country

| | |
|---|---|
| United States | 272 |
| Germany | 267 |
| France | 236 |
| United Kingdom | 230 |
| Russia | 118 |
| Italy | 58 |
| Japan | 57 |
| Austria | 48 |
| Switzerland | 32 |
| Netherlands | 31 |

**C** Illustrious people by language

| | |
|---|---|
| English | 461.4 |
| German | 330.9 |
| French | 258.1 |
| Russian | 118.0 |
| Italian | 61.0 |
| Spanish | 59.9 |
| Japanese | 57.0 |
| Dutch | 47.2 |
| Czech | 28.0 |
| Chinese | 22.1 |

Table 4.8: GLN centrality and number of illustrious people per language according to *Human Accomplishment* (HA). **A** Regression table explaining the number of people (born 1800-1950) of each language group listed in HA as a function of the language group's GDP per capita, population, and eigenvector (EV) centrality in each of GLNs. Cultural production rankings: the **B** countries and **C** languages that contributed the largest number of people to the HA list.

hub language. These alternatives are not mutually exclusive, since the two mechanisms are likely to reinforce each other. Either alternative would highlight the importance of global languages: the position of a language in the network either enhances the visibility of the content produced in it or signals the earlier creation of culturally relevant achievements. Moreover, the results show that the position of a language in the GLN carries information that is not captured by measures of income or population.

# Chapter 5

# Conclusions

In this thesis I used network science to offer a new and precise characterization of a language's global importance. The global language networks (GLNs), mapped from millions of online and printed linguistic expressions, reveal that the world's languages exhibit a hierarchical structure dominated by a central hub, English, and a halo of intermediate hubs, which include other global languages such as German, French, and Spanish. While languages such as Chinese, Arabic and Hindi are immensely popular, I document an important sense in which these languages are more peripheral to the world's network of linguistic influence. For example, the low volume of translations into Arabic, as indicated by our Index Translationum GLN and matched by the peripheral position of Arabic in the Twitter and Wikipedia GLNs, had been identified as an obstacle to the dissemination of outside knowledge into the Arab world [70].

One might argue that the peripheral position of Chinese, Hindi and Arabic in the GLNs stems from biases in the datasets used, such as the underrepresentation of these languages and of some regional languages to which they connect. Indeed, China censors Twitter, Wikipedia and other forms of communication, and many Indians prefer English to Hindi because it is much easier to type. However, the peripheral role of Chinese, Hindi and Arabic in three global forums of recognized importance—Twitter, Wikipedia, and printed book translations—indicates the limited ability of these languages to spread their ideas around the world, at least for the time being, and weakens their claim for global influence. Of course, Chinese, Hindi or Arabic might be connected to languages that are spoken in their respective regions and are not documented in the datasets we used. However, this would still not make them global hubs, since a global language also connects distant languages, and not just local or regional ones.

The substantial hierarchical structure of the three GLNs points to a variety of causal hypothesis and raises questions about the dynamics and effects of globalization. For example, the structure of the GLNs suggests that the world may enjoy the benefits of worldwide communication without either a dedicated international language such as Esperanto, or the hegemony of English—or any other language—as the world's only global language. Assessments of temporal changes in the structure of the GLNs or in their parameters can identify whether English is gaining or losing influence with respect to the languages of rising powers such as India or China. Such changes, as well as the differences between GLNs based on traditional media (printed books) and new media (Twitter), may help to predict a language's likelihood of global importance, marginalization, and, perhaps in the long term, extinction.

GLN centrality can therefore complement current predictions of language processes, which rely mostly on a language's number of speakers [2, 16] and to a lesser extent on geographical and economic properties of the regions in which it is spoken [63]. Is it time to brush up on your Mandarin then? A prediction based on the GLN centrality of languages, as opposed to their number speakers or their economic power, shows that at least in the foreseeable future, enrolling in a Spanish class would be a better use of your time.

# Appendix A

# Language notation

Each of our three datasets uses a different system for identifying language names. For the sake of consistency, I converted the language identifiers to ISO 639-3 identifiers. ISO 639-3 is a code that aims to define three-letter identifiers for all known human languages [61]. For example, English is represented as *eng*, Spanish as *spa*, Modern Greek as *ell* and Ancient Greek as *grc*.

Some languages are *mutually intelligible* or nearly mutually intelligible with others, such as Serbian and Croatian, Indonesian and Malaysian, and the various regional dialects of Arabic. Because of the similarity of mutually intelligible languages I do not consider their speakers as polyglots. Instead, I merged mutually intelligible languages to *macrolanguages* following the ISO 639-3 Macrolanguage Mappings [61]. For example, I merged 29 varieties of Arabic into one Arabic macrolanguage (labelled by the ISO 639-3 identifier *ara*), and Malaysian, Indonesian, and 34 other Bhasa languages into a Malay macrolanguage (*msa*).

Another reason for consolidating languages is that the language detector I used to identify the language of tweets cannot distinguish between the written forms of many mutually intelligible languages, such as Indonesian and Malaysian and Serbian and Croatian. For this reason, I added a couple of merges that are not in the ISO 639-3 macrolanguage mappings: I consolidated Serbian, Croatian, and Bosnian into Serbo-Croatian (*hbs*) even though the latter had been deprecated as a macrolanguage, and merged Tagalog (*tgl*) with Filipino (*fil*) into one Filipino language that uses the identifier *fil*. The full conversion table is available on the SOM page.

Finally, I mapped languages to language families [57] using the hierarchy in Ethnologue [38] complemented by information from articles from the English Wikipedia about the respective languages. I used the standard language family names and identifiers as defined by ISO 639-5 [39].

# Appendix B

# Demographics

## B.1 Population

I retrieved language speaker estimates from the June 14, 2012 version of the Wikipedia statistics page [76]. These estimate include all speakers of a language, native and non-native alike. I converted language names to ISO 639-3 identifiers and merged them into macrolanguages as explained in Appendix A. Refer to Table B.1 for number of speakers for languages in the GLNs.

In general, the number of speakers of a macrolanguage is the sum of speakers of its constituent languages. However, for the macrolanguages listed in Table B.2 I determined that the estimated number of speakers for one of the individual languages that constitute them includes speakers of the other languages, and used that number as the speaker estimate for the entire macrolanguage.

## B.2 Income

The GDP (*gross domestic product*) per capita for a language $l$ measures the average contribution of a single speaker of language $l$ to the world GDP, and is calculated by summing the contributions of speakers of $l$ to the GDP of every country, and dividing the sum by the number of speakers of $l$. A similar method was used by [18]. Given a country $c$, let $G_c$ be the GDP per capita (based on purchasing-power-parity) of that country (retrieved from

| | Language | Code | Speakers (millions) | GDP per capita ($) |
|---|---|---|---|---|
| 1 | **Afrikaans** | afr | 13 | 5,554 |
| 2 | **Albanian** | sqi | 16 | 1,719 |
| 3 | **Arabic** | ara | 530 | 5,027 |
| 4 | **Armenian** | hye | 6 | 3,265 |
| 5 | **Azerbaijani** | aze | 27 | 3,239 |
| 6 | **Bashkir** | bak | 2 | N/A |
| 7 | **Basque** | eus | 1 | 28,815 |
| 8 | **Belarusian** | bel | 6 | 8,772 |
| 9 | **Bengali** | ben | 230 | 2,729 |
| 10 | **Bulgarian** | bul | 12 | 6,750 |
| 11 | **Catalan** | cat | 9 | 27,214 |
| 12 | **Chinese** | zho | 1575 | 8,003 |
| 13 | **Czech** | ces | 12 | 22,952 |
| 14 | **Danish** | dan | 6 | 34,325 |
| 15 | **Dutch** | nld | 27 | 35,089 |
| 16 | **English** | eng | 1500 | 11,943 |
| 17 | **Esperanto** | epo | 1 | N/A |
| 18 | **Estonian** | est | 1.07 | 16,995 |
| 19 | **Filipino** | fil | 90 | 2,583 |
| 20 | **Finnish** | fin | 6 | 30,195 |
| 21 | **French** | fra | 200 | 16,622 |
| 22 | **Galician** | glg | 4 | 25,213 |
| 23 | **Georgian** | kat | 4 | 5,020 |
| 24 | **German** | deu | 185 | 19,535 |
| 25 | **Greek (Modern)** | ell | 15 | 20,746 |
| 26 | **Haitian** | hat | 12 | 2,322 |
| 27 | **Hebrew** | heb | 10 | 18,810 |
| 28 | **Hindi** | hin | 550 | 3,322 |
| 29 | **Hungarian** | hun | 15 | 14,527 |
| 30 | **Icelandic** | isl | 0.32 | 37,250 |
| 31 | **Italian** | ita | 70 | 27,715 |
| 32 | **Japanese** | jpn | 132 | 33,521 |
| 33 | **Kara-Kalpak** | kaa | 0.41 | |
| 34 | **Kazakh** | kaz | 12 | 11,391 |
| 35 | **Kirghiz** | kir | 5 | 1,687 |
| 36 | **Korean** | kor | 78 | 19,866 |
| 37 | **Latin** | lat | | N/A |
| 38 | **Latvian** | lav | 2.15 | 9,292 |

| | Language | Code | Speakers (millions) | GDP per capita ($) |
|---|---|---|---|---|
| 39 | **Lithuanian** | lit | 4 | 13,665 |
| 40 | **Macedonian** | mkd | 3 | 4,785 |
| 41 | **Malay** | msa | 300 | 5,579 |
| 42 | **Malayalam** | mal | 37 | 3,849 |
| 43 | **Maltese** | mlt | 0.37 | 25,406 |
| 44 | **Maori** | mri | 0.157 | N/A |
| 45 | **Marathi** | mar | 90 | 3,462 |
| 46 | **Moldavian** | mol | N/A | N/A |
| 47 | **Mongolian** | mon | 5 | 3,017 |
| 48 | **Norwegian** | nor | 5 | 50,340 |
| 49 | **Occitan** | oci | 2 | N/A |
| 50 | **Persian** | fas | 107 | 7,352 |
| 51 | **Polish** | pol | 43 | 17,921 |
| 52 | **Portuguese** | por | 290 | 9,535 |
| 53 | **Romanian** | ron | 28 | 9,232 |
| 54 | **Russian** | rus | 278 | 9,437 |
| 55 | **Sanskrit** | san | 0.05 | N/A |
| 56 | **Serbo-Croatian** | hbs | 23 | 7,927 |
| 57 | **Sinhala** | sin | 19 | 4,747 |
| 58 | **Slovak** | slk | 7 | 16,428 |
| 59 | **Slovenian** | slv | 2 | 28,593 |
| 60 | **Spanish** | spa | 500 | 13,300 |
| 61 | **Swahili** | swa | 50 | 3,147 |
| 62 | **Swedish** | swe | 10 | 37,727 |
| 63 | **Tajik** | tgk | 4 | 5,045 |
| 64 | **Tamil** | tam | 66 | 4,311 |
| 65 | **Tatar** | tat | 8 | N/A |
| 66 | **Thai** | tha | 73 | 8,636 |
| 67 | **Tibetan** | bod | 7 | |
| 68 | **Turkish** | tur | 70 | 15,156 |
| 69 | **Turkmen** | tuk | 9 | 3,173 |
| 70 | **Uighur** | uig | 10 | N/A |
| 71 | **Ukrainian** | ukr | 45 | 4,841 |
| 72 | **Urdu** | urd | 60 | 4,416 |
| 73 | **Uzbek** | uzb | 24 | 3,125 |
| 74 | **Vietnamese** | vie | 80 | 3,842 |
| 75 | **Welsh** | cym | 0.75 | 36,406 |
| 76 | **Yiddish** | yid | 3 | N/A |

Table B.1: Population and GDP per capita for languages in the GLNs.

| Macrolanguage | ISO 639-3 identifier | Speaker estimate we use in our dataset | Individual languages according to Wikipedia (Wikipedia language code) | Wikipedia Statistics speaker estimate |
|---|---|---|---|---|
| **Akan** | aka | 19 million | Akan (ak) | 19 million |
| | | | Twi (tw) | 15 million |
| **Arabic** | ara | 530 million | Arabic (ar) | 530 million |
| | | | Egyptian Arabic (arz) | 76 million |
| **Malay** | msa | 300 million | Malay (ms) | 300 million |
| | | | Indonesian (id) | 250 million |
| **Serbo-Croatian** | hbs | 23 million | Serbo-Croatian (sh) | 23 million |
| | | | Serbian (sr) | 23 million |
| | | | Croatian (hr) | 6 million |
| | | | Bosnian (bs) | 3 million |
| **Norwegian** | nor | 5 million | Norwegian (no) | 5 million |
| | | | Nynorsk (nn) | 5 million |
| **Komi** | kom | 293,000 | Komi (kv) | 293,000 |
| | | | Komi-Perniak (koi) | 94,000 |

Table B.2: Macrolanguages for which the estimated number of speakers is not an sum of the estimates for the individual languages that constitute them.

[37]) and let $N_c$ be its population, retrieved from [12]. Also, given a language $l$, let $N_{lc}$ be the number of speakers of $l$ in country $c$, obtained from [38] and [12]. I calculated $N_{lc}$ using the language demographics listed in Table 4.2. Thus, $G_l$, the GDP per capita for $l$ is

$$G_l = \frac{\sum_c G_c \frac{N_{lc}}{N_c}}{\sum_c N_{lc}} \tag{B.1}$$

Refer to Table B.1 for GDP per capita for languages in the GLNs.

# Appendix C

# Regression tables for all years

In Section 4.2, I presented the correlation between the centrality of a language and the number of illustrious people associated with this language, as determined by two independent datasets. I considered only people born between 1800 and 1950 to improve the accuracy of the country-to-language mappings. In addition, I removed Albanian from the *Human Accomplishment* regressions, as this language proved to be a major outlier and the number of illustrious people associated with it was negligible (0.05).

This appendix presents the results of the regressions without any restrictions on year of birth, for the *Wikipedia 20* dataset (e.g., people with articles in at least 20 Wikipedia language editions, Table C.1) and the *Human Accomplishment* dataset (Table C.2). In addition, I present here a version of the Human Accomplishment regression table that includes Albanian (Table C.3).

As in Section 4.2, the correlation between the number of illustrious people and the eigenvector centrality of a language in the Wikipedia or book translation networks—though not the Twitter network—is higher than the correlation observed between the number of illustrious people and the income and population of the language. Also, in both the restricted and unrestricted regressions, eigenvector centrality in the Twitter network does not explain much of the variance in number of illustrious people per language.

For full listings of language population and GDP per capita, refer to Table B.1. For full listings of language centrality measures, refer to Table 4.1. For full listings of number of illustrious people by country and language, refer to Tables 4.3 to 4.6.

**A**

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| | Number of illustrious people born per language, based on having biographies in over 20 Wikiepdia language editions | | | | | | |
| $\log_{10}$(Population) | 0.616*** | | | | 0.535*** | 0.045 | 0.281*** |
| | (0.111) | | | | (0.142) | (0.109) | (0.077) |
| $\log_{10}$(GDP per capita) | 1.106*** | | | | 1.006*** | -0.079 | 0.195 |
| | (0.255) | | | | (0.277) | (0.236) | (0.187) |
| EV centrality [Twitter] | | 1.461*** | | | 0.401 | | |
| | | (0.408) | | | (0.436) | | |
| EV centrality [Wikipedia] | | | 2.095*** | | | 2.070*** | |
| | | | (0.180) | | | (0.295) | |
| EV centrality [book trans.] | | | | 2.196*** | | | 1.872*** |
| | | | | (0.208) | | | (0.229) |
| (Intercept) | -3.619** | 1.569*** | 1.302*** | 1.091*** | -3.180* | 1.564 | -0.046 |
| | (1.145) | (0.139) | (0.074) | (0.097) | (1.243) | (1.048) | (0.803) |
| Observations | 38 | 38 | 38 | 38 | 38 | 38 | 38 |
| p-value | 0 | 0.001 | 0 | 0 | 0 | 0 | 0 |
| R-squared | 0.502 | 0.263 | 0.79 | 0.755 | 0.514 | 0.796 | 0.832 |
| Adjusted R-squared | 0.473 | 0.242 | 0.784 | 0.748 | 0.471 | 0.778 | 0.817 |

***,**,* significant at 0.1%, 1% and 5% levels, respectively. Standard errors in parentheses.

**B** **Illustrious people by country**

| | |
|---|---|
| United States | 3726 |
| United Kingdom | 1515 |
| France | 930 |
| Germany | 798 |
| Italy | 644 |
| Russia | 372 |
| Spain | 346 |
| Japan | 216 |
| Netherlands | 195 |
| Canada | 192 |

**C** **Illustrious people by language**

| | |
|---|---|
| English | 5071.0 |
| French | 1145.8 |
| German | 1019.9 |
| Spanish | 994.6 |
| Italian | 661.9 |
| Russian | 449.6 |
| Dutch | 267.5 |
| Portuguese | 264.6 |
| Japanese | 216.0 |
| Arabic | 194.2 |

Table C.1: GLN centrality and the number of illustrious people per language according to *Wikipedia 20*, without restrictions on year of birth. **A** Regression table explaining the number of people of each language about which there are articles in at least 20 Wikipedia language editions as a function of the language's GDP per capita, population, and eigenvector (EV) centrality in each of GLNs. Cultural production rankings: the **B** countries and **C** languages that produced the largest number of people about which there are articles in at least 20 Wikipedia editions.

**A**

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| | Number of illustrious people per language, based on inclusion in *Human Accomplishment* | | | | | | |
| $\log_{10}$(Population) | 1.043*** | | | | 1.042*** | 0.727** | 0.750*** |
| | (0.124) | | | | (0.174) | (0.232) | (0.140) |
| $\log_{10}$(GDP per capita) | 1.866*** | | | | 1.864*** | 1.180* | 1.170** |
| | (0.334) | | | | (0.399) | (0.538) | (0.359) |
| EV centrality [Twitter] | | 2.129*** | | | 0.004 | | |
| | | (0.566) | | | (0.515) | | |
| EV centrality [Wikipedia] | | | 2.293*** | | | 0.882 | |
| | | | (0.322) | | | (0.553) | |
| EV centrality [book trans.] | | | | 2.501*** | | | 1.237** |
| | | | | (0.404) | | | (0.385) |
| (Intercept) | -8.093*** | 0.899*** | 0.617*** | 0.314 | -8.086*** | -5.019* | -5.255** |
| | (1.531) | (0.196) | (0.150) | (0.211) | (1.820) | (2.435) | (1.583) |
| Observations | 29 | 29 | 29 | 29 | 29 | 29 | 29 |
| p-value | 0 | 0.001 | 0 | 0 | 0 | 0 | 0 |
| R-squared | 0.733 | 0.344 | 0.652 | 0.587 | 0.733 | 0.757 | 0.811 |
| Adjusted R-squared | 0.712 | 0.32 | 0.639 | 0.571 | 0.701 | 0.728 | 0.788 |

***,**,* significant at 0.1%, 1% and 5% levels, respectively. Standard errors in parentheses.

**B** **Illustrious people by country**

| | |
|---|---|
| France | 542 |
| Germany | 536 |
| United Kingdom | 531 |
| Italy | 389 |
| United States | 297 |
| China | 237 |
| Japan | 169 |
| Russia | 134 |
| Ancient Greece | 134 |
| India | 93 |

**C** **Illustrious people by language**

| | |
|---|---|
| English | 772.1 |
| German | 645.1 |
| French | 593.6 |
| Italian | 394.8 |
| Chinese | 237.1 |
| Japanese | 169.0 |
| Russian | 134.0 |
| Dutch | 133.2 |
| Spanish | 100.5 |
| Arabic | 86.0 |

Table C.2: GLN centrality and number of illustrious people per language according to *Human Accomplishment* (HA), without any restriction on year of birth. **A** Regression table explaining the number of people of each language listed in HA as a function of the language's GDP per capita, population, and eigenvector (EV) centrality in each of GLNs. Cultural production rankings: the **B** countries and **C** languages that contributed the largest number of people to the HA list.

**A**

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| | Number of illustrious people born 1800-1950 per language, based on inclusion in *Human Accomplishment* | | | | | | |
| $\log_{10}$(Population) | 0.896*** | | | | 0.856*** | 0.538** | 0.535*** |
| | (0.113) | | | | (0.146) | (0.170) | (0.088) |
| $\log_{10}$(GDP per capita) | 1.981*** | | | | 1.929*** | 1.317*** | 1.226*** |
| | (0.263) | | | | (0.292) | (0.347) | (0.197) |
| EV centrality [Twitter] | | 2.064** | | | 0.210 | | |
| | | (0.618) | | | (0.479) | | |
| EV centrality [Wikipedia] | | | 2.474*** | | | 1.174* | |
| | | | (0.308) | | | (0.446) | |
| EV centrality [book trans.] | | | | 2.960*** | | | 1.786*** |
| | | | | (0.323) | | | (0.265) |
| (Intercept) | -8.652*** | 0.523* | 0.180 | -0.269 | -8.428*** | -5.723** | -5.722*** |
| | (1.193) | (0.211) | (0.141) | (0.166) | (1.315) | (1.551) | (0.852) |
| Observations | 30 | 30 | 30 | 30 | 30 | 30 | 30 |
| p-value | 0 | 0.002 | 0 | 0 | 0 | 0 | 0 |
| R-squared | 0.754 | 0.285 | 0.697 | 0.75 | 0.756 | 0.806 | 0.911 |
| Adjusted R-squared | 0.736 | 0.26 | 0.686 | 0.741 | 0.728 | 0.783 | 0.9 |

***,**,* significant at 0.1%, 1% and 5% levels, respectively. Standard errors in parentheses.

**B**

| Illustrious people by country | |
|---|---|
| United States | 272 |
| Germany | 267 |
| France | 236 |
| United Kingdom | 230 |
| Russia | 118 |
| Italy | 58 |
| Japan | 57 |
| Austria | 48 |
| Switzerland | 32 |
| Netherlands | 31 |

**C**

| Illustrious people by language | |
|---|---|
| English | 461.4 |
| German | 330.9 |
| French | 258.1 |
| Russian | 118.0 |
| Italian | 61.0 |
| Spanish | 59.9 |
| Japanese | 57.0 |
| Dutch | 47.2 |
| Czech | 28.0 |
| Chinese | 22.1 |

Table C.3: GLN centrality and the number of illustrious people per language according to *Wikipedia*, for people born 1800-1950, including Albanian. Parts B and C are identical to Table 4.8.

# Bibliography


[1] R. Abramitzky and I. Sin. Book translations as idea flows: The effects of the collapse of communism on the diffusion of knowledge, 2012.

[2] D. M. Abrams and S. H. Strogatz. Linguistics: Modelling the dynamics of language death. *Nature*, 424(6951), 2003.

[3] R. Albert, H. Jeong, and A.-L. Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, 2000.

[4] L. Aronin and D. Singleton. Multilingualism as a new linguistic dispensation. *International Journal of Multilingualism*, 5(1):1–16, 2008.

[5] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[6] A. Blackledge and A. Creese. *Multilingualism: A Critical Perspective*. Continuum, 1 edition, 2010.

[7] P. Bonacich. Power and centrality: A family of measures. *American Journal of Sociology*, pages 1170–1182, 1987.

[8] R. Bouckaert, P. Lemey, M. Dunn, S. J. Greenhill, A. V. Alekseyenko, A. J. Drummond, R. D. Gray, M. A. Suchard, and Q. D. Atkinson. Mapping the origins and expansion of the Indo-European language family. *Science*, 337(6097):957–960, 2012.

[9] d. boyd and K. Crawford. Six provocations for big data. *SSRN eLibrary*, 2011.

[10] C. Brohy. Generic and/or specific advantages of bilingualism in a dynamic plurilingual situation: The case of French as official L3 in the School of Samedan (Switzerland). *International Journal of Bilingual Education and Bilingualism*, 4(1):38–49, 2001.

[11] L. L. Cavalli-Sforza, P. Menozzi, and A. Piazza. *The History and Geography of Human Genes*. Princeton University Press, 1994.

[12] Central Intelligence Agency. *The World Factbook*. Central Intelligence Agency, Washington, DC, 2011.

[13] J. K. Chambers. *Sociolinguistic Theory: Linguistic Variation and its Social Significance*. Wiley-Blackwell, Oxford, rev. ed. edition, 2009.

[14] R. Cohen, K. Erez, D. Ben-Avraham, and S. Havlin. Resilience of the Internet to random breakdowns. *Physical review letters*, 85(21):4626–4628, 2000.

[15] R. Comanaru and K. A. Noels. Self-determination, motivation, and the learning of Chinese as a heritage language. *Canadian Modern Language Review/ La Revue canadienne des langues vivantes*, 66(1):131–158, 2009.

[16] D. Crystal. *Language Death.* Cambridge University Press, Cambridge, UK, Apr. 2000.

[17] D. Crystal. *English as a Global Language*. Cambridge University Press, Cambridge, UK, 2003.

[18] M. Davis. GDP by language. Technical Report 13, Unicode Consortium. http://www.unicode.org/notes/tn13. Published Jan. 22, 2003, accessed Oct. 1, 2012.

[19] M. Erard. *Babel No More: The Search for the World's Most Extraordinary Language Learners*. Free Press, New York, 2012.

[20] European Commission. Eurobarometer report 54: Europeans and languages. executive summary. Technical Report ZA3389, European Opinion Research Group, 2001.

[21] European Commission. Eurobarometer 2001.1: Candidate countries. Technical Report ZA3978, European Opinion Research Group, 2004.

[22] Freebase. person.tsv. http://download.freebase.com/datadumps/latest/browse/people, 2012. Published Nov. 9, 2012, accessed Dec. 20, 2012.

[23] Freebase Wiki. Freebase API. http://wiki.freebase.com/wiki/Freebase_API.

[24] R. C. Gardner. *Motivation and Second Language Acquisition: The Socio-Educational Model.* Peter Lang, Oct. 2010.

[25] R. C. Gardner and W. E. Lambert. Motivational variables in second-language acquisition. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 13(4):266–272, 1959.

[26] Google. The Google Geocoding API v3. https://developers.google.com/maps/documentation/geocoding/.

[27] M. Graham. Wiki space: Palimpsests and the politics of exclusion. In G. W. Lovink and N. Tkacz, editors, *Critical Point of View: A Wikipedia Reader*, pages 269–282. Institute of Network Cultures, 2011.

[28] R. D. Gray and Q. D. Atkinson. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(6965):435–439, 2003.

[29] F. Grin. *Language Policy in Multilingual Switzerland: Overview and Recent Developments*. European Centre for Minority Issues, 1999.

[30] S. A. Hale. Net increase? Cross-Lingual linking in the blogosphere. *Journal of Computer-Mediated Communication*, 17(2):135–151, 2012.

[31] S. A. Hale, D. Gaffney, and M. Graham. Where in the world are you? Geolocation and language identification in Twitter. Technical report, Working paper, 2012.

[32] B. Hecht and D. Gergle. Measuring self-focus bias in community-maintained knowledge repositories. In *Proceedings of the fourth international conference on Communities and technologies*, C&T '09, pages 11–20, New York, NY, USA, 2009. ACM.

[33] J. Heilbron. Towards a sociology of translation: Book translations as a cultural world-system. *European Journal of Social Theory*, 2(4):429–444, 1999.

[34] S. C. Herring, J. C. Paolillo, I. Ramos-Vielba, I. Kouper, E. Wright, S. Stoerger, L. A. Scheidt, and B. Clark. Language networks on LiveJournal. In *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on*, 2007.

[35] R. F. i Cancho and R. V. Solé. The small world of human language. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1482):2261–2265, 2001.

[36] International Information Centre for Terminology. ISO 639-1 registration authortity, 2002.

[37] International Monetary Fund. World Economic Outlook Database, April 2012. Technical report, International Monetary Fund, 2012.

[38] M. P. Lewis. *Ethnologue: Languages of the World.* SIL international, Dallas, TX, 16 edition, 2009.

[39] Library of Congress. ISO 639-5 registration authority, 2008.

[40] S. Lobachev. Top languages in global information production. *Partnership: the Canadian Journal of Library and Information Practice and Research*, 3(2), 2008.

[41] S. Lovgren. English in decline as a first language, study says. *National Geographic News*. Published Feb. 26, 2004, accessed Apr. 26, 2013.

[42] M. McCandless. Chromium Compact Language Detector, May 2011.

[43] Meta-Wiki. List of Wikipedias. http://meta.wikimedia.org/wiki/List_of_Wikipedias.

[44] D. Mocanu, A. Baronchelli, B. Gonçalves, N. Perra, and A. Vespignani. The twitter of babel: Mapping world languages through microblogging platforms. *arXiv:1212.5238*, 2012.

[45] C. A. Murray. *Human Accomplishment: The Pursuit of Excellence in the Arts and Sciences, 800 B.C. to 1950.* HarperCollins, New York, 2003.

[46] N. Ostler. *Empires of the Word: a Language History of the World*. HarperCollins Publishers, New York, 2005.

[47] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: bringing order to the web. Technical report, Stanford InfoLab, 1999.

[48] J. Pak. Is English or Mandarin the language of the future? *BBC*. Published Feb. 21, 2012, accessed Apr. 2, 2013.

[49] Pew Internet & American Life Project. Twitter reaction to events often at odds with overall public opinion. Technical report, Pew Internet & American Life Project. Published Mar. 4, 2013, accessed Mar. 6, 2013.

[50] D. Pimienta, D. Prado, and A. Blanco. *Twelve Years of Measuring Linguistic Diversity in the Internet: Balance and Perspectives*. United Nations Educational, Scientific and Cultural Organization, 2009.

[51] A. Portes and L. Hao. The price of uniformity: language, family and personality adjustment in the immigrant second generation. *Ethnic and Racial Studies*, 25(6):889–912, 2002.

[52] A. Portes and R. G. Rumbaut. *Legacies: The Story of the Immigrant Second Generation*. University of California Press, 2001.

[53] F. Rahimi and M. S. Bagheri. On the status of English as a "lingua franca": An EFL academic context survey. *Studies in Literature & Language*, 3(2):118–122, 2011.

[54] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, 2002.

[55] S. Rodriguez. Another milestone for Twitter: 200 million monthly active users. *Los Angeles Times*. Published Dec. 19, 2012, accessed Mar. 12, 2013.

[56] S. Romaine. The bilingual and multilingual community. In T. J. Bhatia and W. C. Ritchie, editors, *The Handbook of Bilingualism and Multilingualism*, pages 385–405. Blackwell, Malden, MA, 2004.

[57] M. Ruhlen. *A Guide to the World's Languages: Classification*. Stanford University Press, 1991.

[58] H. F. Schiffman. *Linguistic Culture and Language Policy*. Psychology Press, 1998.

[59] B. Seidlhofer. Closing a conceptual gap: The case for a description of English as a lingua franca. *International Journal of Applied Linguistics*, 11(2):133–158, 2001.

[60] B. Seidlhofer, A. Breiteneder, and M.-L. Pitzl. English as lingua franca in europe: Challenges for applied linguistics. *Annual Review of Applied Linguistics*, 26:3–34, 2006.

[61] SIL International. ISO 639-3 registration authortity, 2007.

[62] H. A. Simon. The architecture of complexity. *Proceedings of the American Philosophical Society*, pages 467–482, 1962.

[63] W. J. Sutherland. Parallel extinction risk and global distribution of languages and species. *Nature*, 423(6937):276–279, 2003.

[64] Y. Takhteyev, A. Gruzd, and B. Wellman. Geography of twitter networks. *Social Networks*, 34(1):73–81, 2012.

[65] "Tom". Is teaching english in china a waste of time? *Seeing Red in China*. http://seeingredinchina.com/2011/08/30/is-teaching-english-in-china-a-waste-of-time. Published Aug. 30, 2011, accessed May 8, 2013.

[66] A. Trusina, S. Maslov, P. Minnhagen, and K. Sneppen. Hierarchy measures in complex networks. *Physical Review Letters*, 92(17):178702, 2004.

[67] Twitter. Twitter turns six. http://blog.twitter.com/2012/03/twitter-turns-six.html. Published Mar. 21, 2012, accessed Nov. 14, 2012.

[68] UNESCO. Index Translationum: Contributions from countries. http://www.unesco.org/xtrans/bscontrib.aspx. Accessed July 22, 2012.

[69] UNESCO. Index Translationum: World bibliography of translation. http://www.unesco.org/xtrans/bsform.aspx.

[70] United Nations Development Programme. Arab human development report 2003: Building a knowledge society. Technical report, 2003.

[71] A. Vázquez, R. Pastor-Satorras, and A. Vespignani. Large-scale topological and dynamical properties of the internet. *Physical Review E*, 65(6):066130, 2002.

[72] L. Venuti. *The Translator's Invisibility: A History of Translation*. Routledge, London; New York, 1995.

[73] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *nature*, 393(6684):440–442, 1998.

[74] G. Weber. The world's 10 most influential languages. *Language Today*, 2(3):12–18, 1997.

[75] Wikimedia. MediaWiki API. https://www.mediawiki.org/wiki/API.

[76] E. Zachte. Wikipedia statistics. http://stats.wikimedia.org/EN/Sitemap.htm, 2012. Published June 14, 2012, accessed July 1, 2012.

[77] E. Zuckerman. Meet the bridgebloggers. *Public Choice*, 134(1):47–65, 2008.

[78] E. Zuckerman. *Rewire: Digital Cosmopolitans in the Age of Connection*. WW Norton, New York, 2013.