Links that speak: The global language network and its association with global fame

Citation	Ronen, Shahar, Bruno Gonçalves, Kevin Z. Hu, Alessandro Vespignani, Steven Pinker, and César A. Hidalgo. 2014. "Links That Speak: The Global Language Network and Its Association with Global Fame." Proc Natl Acad Sci USA 111 (52) (December 15): E5616–E5622. doi:10.1073/pnas.1410931111.
Published Version	10.1073/pnas.1410931111
Citable link	http://nrs.harvard.edu/urn-3:HUL.InstRepos:23680428
Terms of Use	This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Open Access Policy Articles, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#OAP

Links that speak: The global language networks of Twitter, Wikipedia and Book Translations

Shahar Ronen[1], Bruno Gonçalves[2,3,4], Kevin Z. Hu[1], Alessandro Vespignani[2], Steven A. Pinker[5], César A. Hidalgo[1]

[1] Macro Connections, The MIT Media Lab, Cambridge, MA 02139, USA
[2] Department of Physics, Northeastern University, Boston, MA 02115, USA
[3] Aix-Marseille Université, CNRS, CPT, UMR 7332, 13288 Marseille, France
[4] Université de Toulon, CNRS, CPT, UMR 7332, 83957 La Garde, France
[5] Department of Psychology, Harvard University, Cambridge, MA 02138, USA

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. We propose that the structure of the network connecting multilingual speakers and translated texts provides a concept of language importance that carries information about the global influence of a language that goes beyond simple economic or demographic measures. We present three independent maps of Global Language Networks (GLN) constructed from millions of records of online and printed linguistic expressions taken from Wikipedia, Twitter, and UNESCO's book translation database. We find that the structure of the three GLNs is centered on English as a global hub, and a handful of intermediate hub languages, which include Spanish, German, French, Russian, Portuguese and Chinese. We validate the measure of a language's centrality in the three GLNs by showing that they exhibit a strong correlation with two independent measures of the number of famous people born in the countries associated with that language. We suggest that other phenomena of a language's present and future influence are systematically related to the structure of the Global Language Network.

Significance Statement

People have long debated about the global influence of languages. Recently, this debate has been fueled by speculations about the relative importance of English and Chinese. These speculations, however, rely on measures of language importance—such as income and population—that lack external validation as measures of a language's global influence. Here we introduce a metric of language influence by characterizing the position of each language in the network connecting languages that are co-spoken. We show that the connectivity of a language in this network, after controlling for the number of speakers of a language and their income, remains a strong predictor of a language's influence when validated against two independent measures of the cultural content produced by a language's speakers.

Introduction

Of the thousands of languages that have ever been spoken only a handful have become influential enough to be considered *global languages*. 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[1–4]. Yet demographic and economic measures are unable to capture an important aspect of the global influence of a language[5]: its ability to intermediate information among speakers that do not speak that language.

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. Here, the widespread use of a standard language for a specific purpose forces people in a certain profession to acquire it, making it even more widespread[6]. Major conquests, such as those undertaken by the Roman Empire and colonial Europe, 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[7], 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 *lingua franca* that facilitates global communication and protecting the local languages that strengthen cultural diversity and ethnic or national pride. 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.

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[9].

Despite the importance of global languages, there is no rigorous formulation of the concept of a global language, nor a way to measure the degree to which a language is global. In this paper we use network science to develop a metric for measuring the global influence of languages and to define what a global language is. Our 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[10].

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.

Several languages act as *hub languages* that provide indirect paths between the speakers of many other languages. We distinguish between two kinds of hubs: *global hubs* that provide indirect links between distant languages, such as Portuguese and Vietnamese, and r*egional hubs* that provide indirect links between local languages or regional dialects. For instance, the connections that Chinese provide between Tibetan, Mongolian and Uighur represent connections between languages that are spoken primarily in China, thus making Chinese a regional hub—rather than global—in this example.

Finally, we note that our network connects languages that share speakers and literary content. As such, it offers a different perspective from the phylogenetic trees that connect languages with similar origins[11], or the semantic networks that connect synonyms or words co-occurring frequently in text[12].

Data and Methods

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. Accordingly, we 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[13].

Going forward, we note that **the resulting networks represent patterns of linguistic co-expression, not among the entire human population, but 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).** Additional datasets could be used to map the language networks of other groups as long as these datasets cover the linguistic expression of a large fraction of multilingual speakers. To that extent, monolingual resources, such as the Chinese microblogging service *Sina Weibo*, the Russian social network *VK* or the Chinese encyclopedia *Baidu Baike*, do not represent resources that can be used to map connections between global languages.

The elites that participate of Twitter, Wikipedia and book translations are not representative of the entire human population, yet still represent groups that are worthy of study since elites often drive the cultural, political, technological, and economic processes with which observers of global language patterns are concerned.

Constructing three Global Language Networks

To construct the three global language networks we need to first identify the links that are statistically significant with respect to the population of speakers expressed in each of the three datasets. A statistical significant connection is a connection where the probability of finding a speaker, or record, connecting languages *i* and *j* is larger than what we would expect based on the prevalence of these languages alone (P(*i*,*j*) > P(*i*)P(*j*)). Here, we use the ϕ-correlation and *t*-statistics to assess the statistical significance of language connections.

Let M_{ij} be the matrix representing the number of users, or translations, from language *i* to language *j*. Then the correlation ϕ_{ij} between languages *i* and *j* is given by:

$$\phi_{ij} = \frac{M_{ij}N - M_i M_j}{\sqrt{M_i M_j (N - M_i)(N - M_j)}} \qquad (1)$$

where N_i represents the number of multilingual users (or translations) expressed in language i $(N_i = \sum_j M_{ij})$ and *N* represents the total number of users or translations in the dataset. ϕ_{ij} is positive for languages than tend to co-occur more than what we would expect based on their representation in a dataset, and negative otherwise. To assess the statistical significance of these correlations we use the *t*-statistic, which is given by:

$$t_{ij} = \frac{\phi_{ij}\sqrt{D-2}}{\sqrt{1 - {\phi_{ij}}^2}} \qquad (2)$$

where $D - 2$ represents the degrees of freedom of the correlation. Here we consider $D = \max(N_i, N_j)$ since this provides a more stringent criteria for finding a correlation than using D=N.

Finally, we construct our network by consider only links that are statistically significant with a p-value < 0.01 (t_{ij} > 2.59 for D>20 (one-tailed)). Also, we consider only links for which M_{ij}>5 to avoid the false positives that could emerge due to small statistics. In sum, we set M_{ij} = 0 if either t_{ij}< 2.59 or M_{ij} < 6. Finally we note that, by definition, a null model network would contain no links since none of the links of a null model network would satisfy the statistical significance condition. Also, we note that, unlike the Twitter and Wikipedia GLN, the book translation GLN is directed, since in this case we have information about the source language of the translation.

Data

Next, we describe the basic properties of the three datasets used (details can be found in the supplementary material). We compiled our Twitter dataset from over one billion tweets collected between December 6, 2011 and February 13, 2012. The language of each tweet was detected using the Chromium Compact Language Detector[14], after removing misleading expressions, such as URLs, hashtags, and @-mentions. We used only tweets that the language detector identified with a certainty score higher than 90% (see SM). Our final dataset consists of nearly 550 million tweets in 73 languages generated by over 17 million unique users, which represented over 10% of Twitter's active users at the time the data were collected. The dataset allows us to estimate the conditional probability that a user tweets in one language, given that he or she tweeted in another language (eqn. (1)).

The Wikipedia dataset was compiled from the edit histories of all Wikipedia language editions as recorded by the end of 2011. After removing edits made by Wikipedia's maintenance bots and applying the filters described in the supplementary material, the dataset contains 382 million edits in 238 languages by 2.5 million unique editors. Here, two languages are connected according to the probability that a user who edits an article in one Wikipedia language edition edits an article in another (eqn. 1). This allows us to approximate the probability that digitally engaged knowledge specialists speak a pair of languages with a high level of mastery.

Finally, the Index Translationum (IT) dataset consists of 2.2 million translations of printed books published between 1979 and 2011 in 150 countries and over a thousand languages (see SM). The dataset records translations rather than books, so it does not list books that have not been translated. Moreover, the IT dataset 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 we treat each independent translation separately, and in this case, count 22 translations from Russian to English. Also we 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). Yet only 13 of these were translated directly from English, and the other were translated from Spanish and Galician. This characteristic of the dataset allows us to identify languages that serve as intermediaries for translations. We use eqn. (2) to estimate the probability that a book was translated into a language from another.

In all three cases we collapsed mutually intelligible languages following the ISO 639-3 standard[15]. For example, Indonesian and Malaysian were both coded as "Malay", and the regional dialects of Arabic are all coded as "Arabic". Further Information on data preparation procedures can be found in the SM.

Results

The Structure of Three Global Language Networks

To understand the relative importance of each language, we begin by visualizing the three GLNs (Figure 1). In this visualization, each node represents a language. Node sizes are proportional to the number of speakers of each language—native and non-native[16]. Node colors indicate language families and link colors show the significance of the link (according to its t-statistic). Finally, link widths show the total co-occurrence (M_{ij}).

Even though the three GLNs capture information about the linguistic expression of different communities, they share a number of features. First, the representation of each language across datasets—number of Twitter users, Wikipedia editors, or translations from a language—correlates strongly across the three networks (Figures 2 A-C). Moreover, the co-expressions (number of common twitter users, Wikipedia editors, and average number of book translations from and to a language) are also positively correlated across the three datasets (Figures 2 D-F). This means that a language with a high or low co-expression to another language in one GLN is likely to have a high or low co-expression with that same language in the other GLNs.

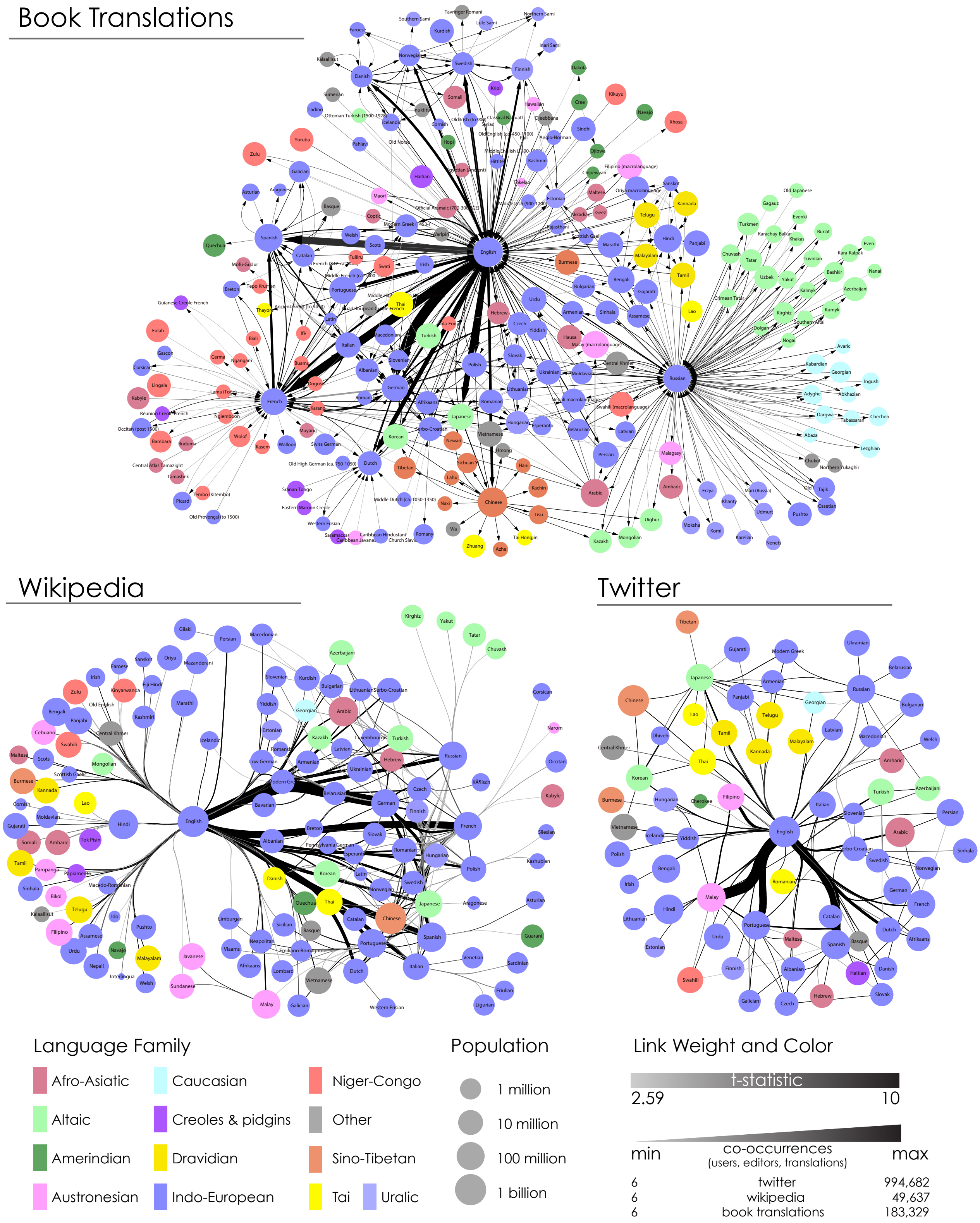


Figure 1 The structures of three Global Language Networks (GLN). The three GLNs contain all language connections that involve at least five users (Twitter and Wikipedia) or five translations, and that are significant with a p-value of less than 0.01.

Interestingly, Wikipedia appears to take the middle ground between Twitter and book translations. The Wikipedia GLN is relatively similar to both the Twitter and Book Translations GLNs, while the latter two are relatively different from each other. Respectively, the R^2 between the representation of a language in the Wikipedia GLN, and in the Book Translations and Twitter GLN, are 63% and 66%. For co-expressions the equivalent numbers are 41% and 63%. By contrast, the similarity between the representation of languages in Twitter and Book Translations is R^2=40%, and their similarity in co-expression is only R^2=23%. Finally, we note that—with respect to the book translation dataset—the two digital datasets (Twitter and Wikipedia) are overexpressed in languages associated with developing countries, like Malay, Filipino and Swahili. This could indicate that these digital media are more inclusive of the populations of developing countries than written books.

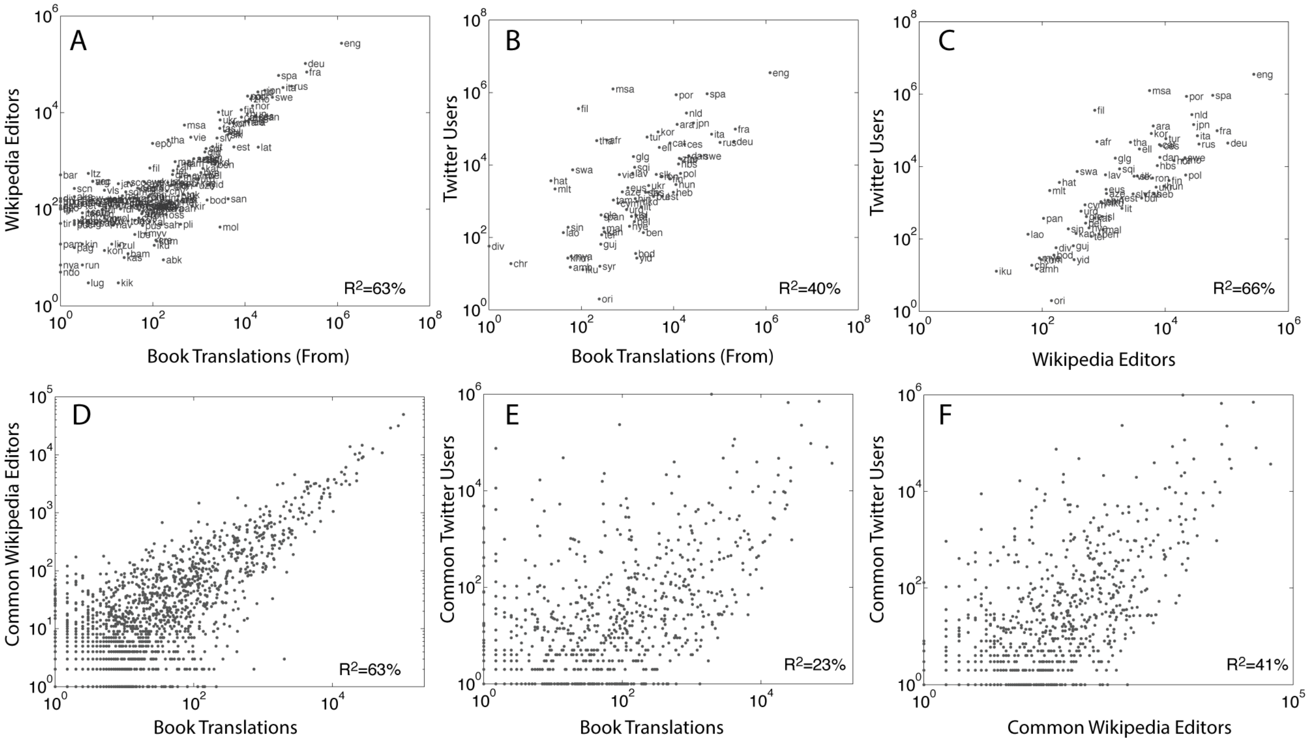


Figure 2 Similarity of the three independent datasets we use for mapping the Global Language Networks. The top row shows the correlation between the number of expressions for each languages across the three datasets: **A** Wikipedia editors in a language and book translations from a language **B** Twitter users in a language and book translations from a language **C** Twitter users and Wikipedia editors. The bottom row shows the correlation between the number of co-expressions for language pairs across different datasets **D** common Wikipedia editors and book translations **E** Common Twitter users and book translations, and **F** common Twitter users and common Wikipedia editors. In D-E we symmetrized the book translation network by considering the average of translations from and to a language.

Next, we study the relationship between the position of a language in the GLN and the global cultural influence of its speakers. We measure the position of a language in the GLN using its eigenvector centrality[26], which is also the basis for Google's PageRank algorithm (for other centrality measures see SM). 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 hubs.

To measure the global influence of the speakers of a language we use two datasets that estimate the number of famous people associated with each language. First, we compiled a list of the 4,886 biographies of people who were born between 1800 and 1950 and have articles in at least 26 Wikipedia language editions (this data is available at pantheon.media.mit.edu). This list is populated by famous individuals of the arts and sciences, such as Einstein, Darwin, Van Gogh and Picasso, by popular writers such as Charles Dickens, social activists such as Che Guevara, as well as by politicians, sportsmen and entrepreneurs. We associated each person with a language using the current language demographics for his or her country of birth. Each famous person in the dataset equals one point, which is distributed across the languages spoken in his or her native country according to its language demographics. For example, a person born in Canada contributes 0.59 to English and 0.22 to French. See SM for a detailed explanation of the conversion and data sources.

The second measure of famous people is based on *Human Accomplishment*, a published volume listing 3,869 individuals that have made significant contributions to the arts and sciences before 1950[27]. We distributed the contribution of the 1,655 people on this list born between 1800-1950 across different languages using the same method used for the Wikipedia dataset (see SM).

Figure S5 (See SM) compares these two independent measures of fame by looking at the correlation between the scores reported in Human Accomplishment and the number of different language editions in which a biography is present in Wikipedia. The correlation between both datasets is mild but significant (R^2=0.25, p-value<<0.001). The mild correlation between the two datasets highlights the robustness of results that hold for both datasets: the differences between the two datasets imply that a result obtained for one does not need to hold for the other merely because of the co-linearity of the data.

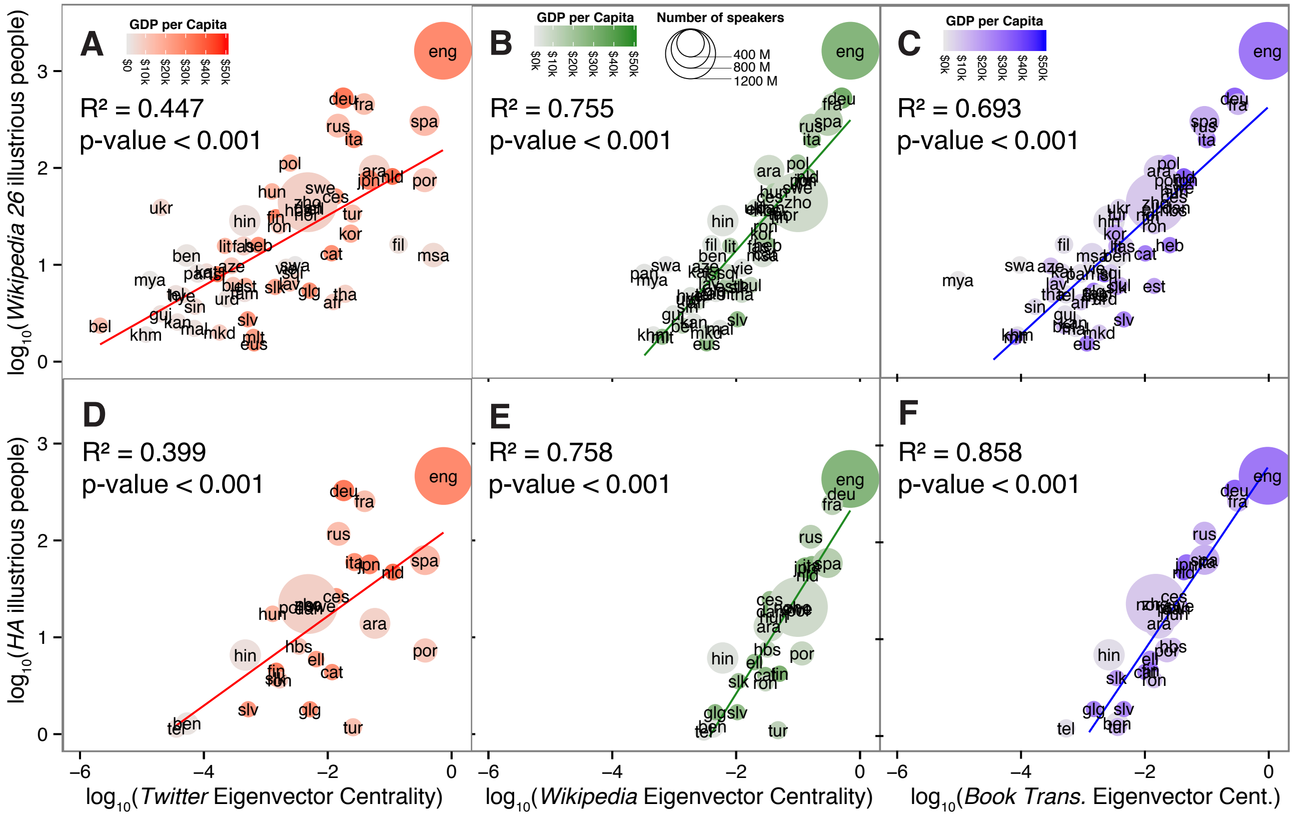


Figure 3 The position of a language in the GLN and the global impact of its speakers. Top row shows the number of people per language (born 1800-1950) with articles in at least 26 Wikipedia language editions as a function of their language's eigenvector centrality in the **A** Twitter GLN, **B** Wikipedia GLN, and **C** book translation GLN. The bottom row shows the number of people per language (born 1800-1950) listed in *Human Accomplishment* as a function of their language's eigenvector centrality in **D** Twitter GLN, **E** Wikipedia GLN, and **F** book translation GLN. Size represents the number of speakers for each language, and color intensity represents GDP per capita for the language group. All subplots report the adjusted R^2.

Figures 3 A-C show the bivariate correlation between the number of famous people measured using the Wikipedia dataset and the eigenvector centrality of that language in the Twitter, Wikipedia and book translation networks. We only use languages that are present in all three GLNs and that have are associated with one or more famous character. Table 1 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 famous people and the eigenvector centrality of a language is higher than the correlation observed between the number of famous 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 famous people explained by the Wikipedia and book translation GLNs are respectively 7.5% (F=22.97, p-value<0.001) and 7.7% (F=23.48, p-value<0.001) after the effects of income and population have been taken into account. In contrast, the semi-partial contributions of income and population are 5.1% (F=7.74, p-value<0.001) when measured against the Wikipedia GLN, and 11.3% (F=17.32, p-value<0.001) when measured against the book translation GLN.

Figures 3 D-F and Table 2 show the same analysis for *Human Accomplishment*. 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 2), which accounts for 89% of the variance. Centrality in the Wikipedia GLN or book translation GLN alone explains 76% and 86% of the variance, respectively, and 6.1% (F=7.59, p-value=0.01) and 16.1% (F=37.98, p-value<0.001) at the margin, as measured by the semi-partial correlation. The semi-partial contribution of income and population in this case is much lower, being only 2.3% (F=1.43, p-value=0.26) and 2.7% (F=3.13, p-value=0.06) when measured, respectively, against the Wikipedia GLN and book translation GLN.

A

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Number of illustrious people born 1800-1950 per language, based on having biographies in at least 26 Wikiepdia language editions						
$\log_{10}$(Population)	0.669***				0.615***	0.254*	0.397***
	(0.060)				(0.080)	(0.102)	(0.077)
$\log_{10}$(GDP per capita)	1.156***				1.041***	0.138	0.400*
	(0.120)				(0.166)	(0.238)	(0.188)
EV centrality [Twitter]		0.362***			0.055		
		(0.051)			(0.054)		
EV centrality [Wikipedia]			0.731***			0.583***	
			(0.054)			(0.123)	
EV centrality [book trans.]				0.588***			0.376***
				(0.050)			(0.078)
(Intercept)	-4.450***	2.240***	2.626***	2.651***	-3.746***	1.415	-0.064
	(0.529)	(0.158)	(0.112)	(0.132)	(0.872)	(1.315)	(1.018)
Observations	61	61	61	61	61	61	61
p-value	0	0	0	0	0	0	0
R-squared	0.734	0.457	0.759	0.698	0.739	0.81	0.811
Adjusted R-squared	0.725	0.447	0.755	0.693	0.725	0.8	0.801

***,**,* significant at 0.1%, 1% and 5% levels, respectively. Standard errors in parentheses.
Only languages with at least one illustrious person are included.

B

Illustrious people by country	
United States	1221
United Kingdom	508
Germany	407
France	397
Russia	240
Italy	194
Poland	114
Austria	91
Spain	77
Japan	75

C

Illustrious people by language	
English	1617.8
German	524.1
French	455.5
Spanish	305.5
Russian	272.9
Italian	198.1
Polish	112.6
Arabic	94.5
Dutch	81.3
Japanese	75.0

D

Twitter EV Cent.	
English	0.69
Malay	0.49
Portuguese	0.35
Spanish	0.35
Filipino	0.13
Dutch	0.11
Arabic	0.05

E

Wikipedia EV Cent.	
English	0.66
German	0.48
French	0.34
Spanish	0.29
Italian	0.16
Russian	0.15
Dutch	0.13

F

Book translation EV Cent.	
English	0.90
French	0.30
German	0.26
Italian	0.09
Russian	0.09
Spanish	0.09
Japanese	0.04

Table 1 GLN centrality and the number of famous people per language according to Wikipedia. **A** Regression table explaining the number of people (born 1800-1950) of each language group about which there are articles in at least 26 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** language groups that produced the largest number of people about which there are articles in at least 26 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. See SM for the full lists.

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)	0.782***				0.943***	0.321	0.269*
	(0.106)				(0.172)	(0.195)	(0.109)
$\log_{10}$(GDP per capita)	1.862***				2.292***	0.679	0.545
	(0.259)				(0.443)	(0.496)	(0.275)
EV centrality [Twitter]		0.462***			-0.159		
		(0.104)			(0.133)		
EV centrality [Wikipedia]			1.026***			0.678*	
			(0.109)			(0.251)	
EV centrality [book trans.]				0.948***			0.722***
				(0.073)			(0.119)
(Intercept)	-8.122***	2.158***	2.528***	2.798***	-10.573***	-1.381	-0.379
	(1.212)	(0.248)	(0.160)	(0.135)	(2.385)	(2.721)	(1.504)
Observations	29	29	29	29	29	29	29
p-value	0	0	0	0	0	0	0
R-squared	0.728	0.42	0.767	0.863	0.743	0.79	0.89
Adjusted R-squared	0.707	0.399	0.758	0.858	0.712	0.764	0.876

***,**,* significant at 0.1%, 1% and 5% levels, respectively. Standard errors in parentheses.
Only languages with at least one illustrious person are included.

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	466.3
German	329.9
French	255.7
Russian	118.0
Spanish	63.0
Italian	60.1
Japanese	57.0
Dutch	47.2
Czech	26.7
Chinese	22.2

Table 2 GLN centrality and number of famous people per language according to *Human Accomplishment*. **A** Regression table explaining the number of people (born 1800-1950) of each language group listed in *Human Accomplishment* (*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** language groups that contributed the largest number of people to the HA list. For the full lists, see SM.

Finally, we note that 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, or 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 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.

Discussion

In this paper we used network science to offer a new and precise characterization of a language's global importance. The Global Language Networks, 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; we 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[28].

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. However, while these languages may be central in other media, their peripheral role in three global forums of recognized importance—Twitter, Wikipedia, and printed book translations—weakens their claim for global influence. Moreover, Chinese, Hindi or Arabic would not qualify as global hubs even if their connections to regional languages were better documented in our datasets, since a global language also links distant languages, and not just local or regional ones.

The structure of the three Global Language Networks documented here raise important 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 and Wikipedia), may help 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[29].

** The datasets used in this paper are available on our supporting online material (SOM) page: macro.media.mit.edu/gln.*

References

1. Davis, M. *GDP by Language*. (Unicode Consortium, 2003). at <http://www.unicode.org/notes/tn13>
2. Ostler, N. *Empires of the Word: a Language History of the World*. (HarperCollins Publishers, 2005).
3. Pimienta, D., Prado, D. & Blanco, Á. *Twelve Years of Measuring Linguistic Diversity in the Internet: Balance and Perspectives*. (United Nations Educational, Scientific and Cultural Organization, 2009). at <http://www.ifap.ru/pr/2010/n100305c.pdf>
4. Weber, G. The World's 10 Most Influential Languages. *Language Today* **2,** 12–18 (1997).
5. Crystal, D. *English as a Global Language*. (Cambridge University Press, 2003).
6. De Swaan, A. *Words of the world: the global language system*. (Polity, 2001).
7. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes*. (Princeton University Press, 1994).
8. Bouckaert, R. *et al.* Mapping the Origins and Expansion of the Indo-European Language Family. *Science* **337,** 957–960 (2012).

9. Zuckerman, E. *Rewire: Digital Cosmopolitans in the Age of Connection*. (WW Norton, 2013).

10. Chambers, J. K. *Sociolinguistic Theory: Linguistic Variation and its Social Significance*. (Wiley-Blackwell, 2009).

11. Gray, R. D. & Atkinson, Q. D. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426,** 435–439 (2003).

12. I Cancho, R. F. & Solé, R. V. The small world of human language. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **268,** 2261–2265 (2001).

13. UNESCO. Index Translationum: World Bibliography of Translation. at <http://www.unesco.org/xtrans/bsform.aspx>

14. McCandless, M. *Chromium Compact Language Detector*. (2011). at <http://code.google.com/p/chromium-compact-language-detector/>

15. SIL International. ISO 639-3 Registration Authortity. (2007). at <http://www.sil.org/iso639-3>

16. Zachte, E. WIkipedia Statistics. *Wikimedia Statistics* (2012). at <http://stats.wikimedia.org/EN/Sitemap.htm>

17. Barabási, A.-L. & Albert, R. Emergence of Scaling in Random Networks. *Science* **286,** 509–512 (1999).

18. Simon, H. A. The Architecture of Complexity. *Proceedings of the American Philosophical Society* 467–482 (1962).

19. Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabási, A. L. Hierarchical Organization of Modularity in Metabolic Networks. *Science* **297,** 1551–1555 (2002).

20. Vázquez, A., Pastor-Satorras, R. & Vespignani, A. Large-scale Topological and Dynamical Properties of the Internet. *Physical Review E* **65,** 066130 (2002).

21. Heilbron, J. Towards a Sociology of Translation: Book Translations as a Cultural World-System. *European Journal of Social Theory* **2,** 429–444 (1999).

22. Venuti, L. *The Translator's Invisibility: A History of Translation*. (Routledge, 1995).

23. Trusina, A., Maslov, S., Minnhagen, P. & Sneppen, K. Hierarchy Measures in Complex Networks. *Physical Review Letters* **92,** 178702 (2004).

24. Cohen, R., Erez, K., Ben-Avraham, D. & Havlin, S. Resilience of the Internet to Random Breakdowns. *Physical review letters* **85,** 4626–4628 (2000).

25. Albert, R., Jeong, H. & Barabási, A. L. Error and Attack Tolerance of Complex Networks. *Nature* **406,** 378–382 (2000).

26. Bonacich, P. Power and Centrality: A Family of Measures. *American Journal of Sociology* 1170–1182 (1987).

27. Murray, C. A. *Human Accomplishment: The Pursuit of Excellence in the Arts and Sciences, 800 B.C. to 1950*. (HarperCollins, 2003).

28. United Nations Development Programme. *Arab Human Development Report 2003: Building a Knowledge Society*. (2003). at <http://www.arab-hdr.org/publications/other/ahdr/ahdr2003e.pdf>

29. Abrams, D. M. & Strogatz, S. H. Linguistics: Modelling the dynamics of language death. *Nature* **424,** 900–900 (2003).

30. Rodriguez, S. Another Milestone for Twitter: 200 Million Monthly Active Users. *Los Angeles Times* (2012). at <http://www.latimes.com/business/technology/la-fi-tn-twitter-200-million-monthly-active-users-20121219,0,3316419.story>

31. Boyd, D. & Crawford, K. Six Provocations for Big Data. *SSRN eLibrary* (2011). at <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431>

32. Pew Internet & American Life Project. *Twitter Reaction to Events Often at Odds with Overall Public Opinion*. (Pew Internet & American Life Project, 2013). at <http://www.pewresearch.org/2013/03/04/twitter-reaction-to-events-often-at-odds-with-overall-public-opinion/>

33. Mocanu, D., Baronchelli, A., Gonçalves, B., Perra, N. & Vespignani, A. The Twitter of Babel: Mapping World Languages through Microblogging Platforms. *arXiv:1212.5238* (2012). at <http://arxiv.org/abs/1212.5238>

34. Graham, M., Hale, S. A. & Gaffney, D. Where in the world are you? Geolocation and language identification in Twitter. *Professional Geographer* (2013).

35. Herring, S. C. *et al.* Language networks on LiveJournal. in *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on* (2007). at <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4076532>

36. International Information Centre for Terminology. ISO 639-1 Registration Authortity. (2002). at <http://www.infoterm.info/standardization/iso_639_1_2002.php>

37. Erard, M. *Babel No More: The Search for the World's Most Extraordinary Language Learners*. (Free Press, 2012).

38. Meta-Wiki. List of Wikipedias. at <http://meta.wikimedia.org/wiki/List_of_Wikipedias>

39. UNESCO. Contributions from Countries. *Index Translationum* at <http://www.unesco.org/xtrans/bscontrib.aspx>

40. Ruhlen, M. *A Guide to the World's Languages: Classification*. (Stanford University Press, 1991).

41. Lewis, M. P. *Ethnologue: Languages of the World*. (SIL international, 2009). at <http://www.ethnologue.com/16>

42. Library of Congress. ISO 639-5 Registration Authority. (2008). at <http://www.loc.gov/standards/iso639-5>

43. International Monetary Fund. *World Economic Outlook Database, April 2012*. (2012). at <http://www.imf.org/external/pubs/ft/weo/2012/01/weodata/index.aspx>

44. Central Intelligence Agency. *The World Factbook*. (Central Intelligence Agency, 2011).

45. Freeman, L. C. A set of measures of centrality based on betweenness. *Sociometry* 35–41 (1977).

46. Graham, M. in *Critical Point of View: A Wikipedia Reader* (Lovink, G. W. & Tkacz, N.) 269–282 (2011).

47. Hecht, B. & Gergle, D. Measuring Self-focus Bias in Community-maintained Knowledge Repositories. in *Proceedings of the fourth international conference on Communities and technologies* 11–20 (ACM, 2009). doi:10.1145/1556460.1556463

48. Freebase. person.tsv. (2012). at <http://download.freebase.com/datadumps/latest/browse/people/person.tsv>

49. Freebase Wiki. Freebase API. at <http://wiki.freebase.com/wiki/Freebase_API>

50. Wikimedia. MediaWiki API. at <https://www.mediawiki.org/wiki/API>

51. Google. The Google Geocoding API v3. at <https://developers.google.com/maps/documentation/geocoding/>

Author Contributions

S.R. and C.A.H. conceived and led the project. S.R. retrieved and analyzed data and wrote the manuscript. B.G. retrieved the Twitter and Wikipedia data and edited the paper. K.Z.H retrieved and analyzed data. A.V. shared his expertise in theoretical network science and edited the manuscript. S.A.P. shared his expertise in language and linguistics and wrote the manuscript. C.A.H wrote the manuscript. S.R. and C.A.H designed the figures.

Supplementary Materials

S1 Data

S2 Language notation and demographics

S3 Additional Calculations

S4 Language centrality: alternatives and robustness

S5 Famous people per language

* The datasets used in this paper are available on our supporting online material (SOM) page: macro.media.mit.edu/gln.

S1 Data

S1.1 Twitter

Twitter 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 from all over the world, tweeting in many different languages. Of these, 200 million users were active every month[30].

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[31]. For example, Twitter users in the United States are younger and hold more liberal opinions than the general public[32].

We collected 1,009,054,492 tweets between December 6, 2011 and February 13, 2012, through the Twitter *garden hose*, which gives access to 10% of all tweets. We detected the language of each tweet using the Chromium Compact Language Detector (CLD)[14], which was chosen for its wide language support and its relatively accurate detection of short messages[33,34]. However, any automated language detection is prone to errors[35], all the more so when performed on short, informal texts such as tweets. To reduce the effect of such errors, we applied the following methods.

Firstly, to improve detection, we 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, we 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*. We 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 S1A**).

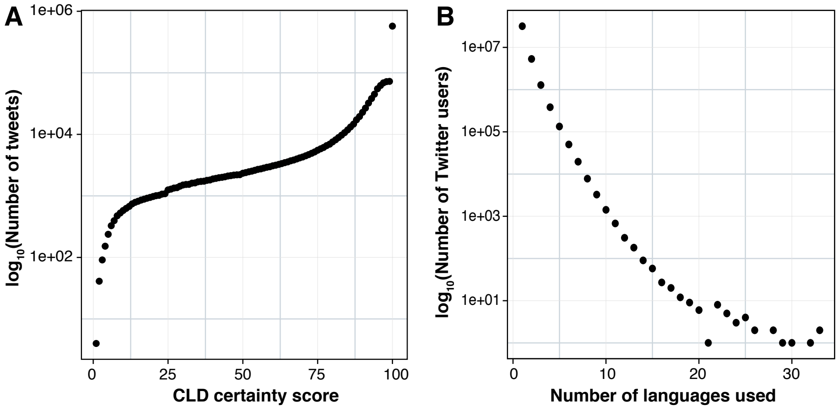


Figure S1 A number of tweets as function of certainty **B** Distribution of Twitter users by number of languages in which they tweet.

Thirdly, as mutually intelligible languages are difficult to distinguish, we merged similar languages. To do so, we converted the two-letter ISO 639-1 language codes[36] produced by CLD to three-letter ISO 639-3 codes[15], and merged them using the ISO 639-3 macrolanguages standard. See Section S2.1 for further details and limitations.

Finally, to reduce the effect of individual detection errors, we 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. We found that a large number of users tweeted in a relatively large number of languages, and we attribute some of this to inaccurate language detection. To prevent this from skewing the representation of the Twitter GLN, we discarded users who tweeted in more than five languages (**Figure S1B**). 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[37]. 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.

After applying the criteria listed above, we had a dataset of 548,285,896 tweets in 73 languages by 17,694,811 users, which is available on the SOM page. We used this dataset to generate the Wikipedia GLN shown in **Figure 1** of the main section. **Table S1** shows statistics for the languages with the most tweets in our Twitter dataset.

#	Language	Code	Tweets	Users	Tweets per user	% of total users
1	**English**	eng	255,351,176	10,859,465	23.5	61.37%
2	**Japanese**	jpn	91,669,691	2,602,426	35.2	14.71%
3	**Malay**	msa	49,546,710	1,651,705	30	9.33%
4	**Portuguese**	por	46,520,572	1,617,409	28.8	9.14%
5	**Spanish**	spa	44,195,979	2,043,468	21.6	11.55%
6	**Korean**	kor	11,674,755	289,982	40.3	1.64%
7	**Dutch**	nld	10,526,980	435,128	24.2	2.46%
8	**Arabic**	ara	9,993,172	366,643	27.3	2.07%
9	**Thai**	tha	7,449,790	154,171	48.3	0.87%
10	**Turkish**	tur	4,660,694	233,158	20	1.32%
11	**Russian**	rus	4,577,942	243,159	18.8	1.37%
12	**French**	fra	3,434,065	147,843	23.2	0.84%
13	**Filipino**	fil	1,905,619	257,611	7.4	1.46%
14	**German**	deu	1,705,256	73,897	23.1	0.42%
15	**Italian**	ita	1,586,225	89,242	17.8	0.50%
16	**Swedish**	swe	596,130	36,604	16.3	0.21%
17	**Modern Greek**	ell	526,527	30,609	17.2	0.17%
18	**Chinese**	zho	453,837	24,113	18.8	0.14%
19	**Catalan**	cat	236,424	32,376	7.3	0.18%
20	**Norwegian**	nor	170,430	16,500	10.3	0.09%

Table S1 Statistics for the twenty languages with the most tweets in our Twitter dataset. The full table is available in the SOM.

S1.2 Wikipedia

Wikipedia 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[38]. Wikipedia's single sign-on mechanism lets editors use the same username on all language editions to which they contribute. This allows us to associate a contribution with a specific person and identify the languages spoken by that person.

We compiled our Wikipedia dataset as follows. Firstly, we collected information on editors and their contributions in different languages from the edit logs of all Wikipedia editions until the end of 2011. We collected 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, we merged the languages as we did for the Twitter dataset, discarding ten Wikipedia editions in the process. Two of them are more or less duplicates of other editions, namely *simple* (Simple English) of English and *be-x-old* (Classic Belarusian) of Official Belarusian. The remaining eight could not be mapped to standard ISO639-3 languages: *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, which may be cosmetic or technical and may not indicate knowledge of a language, we set the same thresholds as for our Twitter dataset. For each user we considered only languages in which he or she made at least two edits, and considered only users who made at least five edits overall. We also discarded editors who contributed to more than five languages, following the rationale explained in the Twitter section. We 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 S2**). For example, one of the users we identified was 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, this user 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.

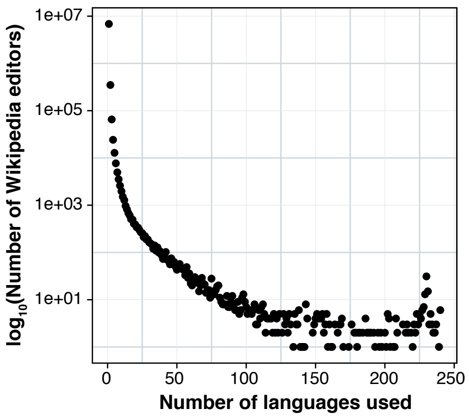


Figure S2 Distribution of Wikipedia editors by number of languages in which they contribute.

Table S2 below shows statistics for the languages with the most edits in our dataset. The final dataset consists of 382,884,184 edits in 238 languages by 2,562,860 contributors, and is available on the SOM page. We used this dataset to generate the Wikipedia GLN shown in **Figure 1** of the main section.

#	Language	Code	Edits	Editors	Edits per user	% of total editors
1	**English**	eng	198,361,048	1,589,250	124.81	62.011%
2	**German**	deu	33,977,378	224,215	151.54	8.749%
3	**French**	fra	23,070,757	142,795	161.57	5.572%
4	**Japanese**	jpn	16,149,315	102,857	157.01	4.013%
5	**Spanish**	spa	13,645,596	145,487	93.79	5.677%
6	**Russian**	rus	12,445,887	81,925	151.92	3.197%
7	**Italian**	ita	11,923,658	72,981	163.38	2.848%
8	**Chinese**	zho	7,302,770	50,341	145.07	1.964%
9	**Polish**	pol	6,589,015	47,015	140.15	1.834%
10	**Dutch**	nld	6,393,791	46,951	136.18	1.832%
11	**Hebrew**	heb	5,467,149	18,998	287.77	0.741%
12	**Portuguese**	por	5,168,734	60,487	85.45	2.360%
13	**Swedish**	swe	3,521,224	30,498	115.46	1.190%
14	**Finnish**	fin	2,926,115	20,811	140.60	0.812%
15	**Hungarian**	hun	2,713,725	18,033	150.49	0.704%
16	**Korean**	kor	2,634,092	16,464	159.99	0.642%
17	**Arabic**	ara	2,178,719	18,258	119.33	0.712%
18	**Turkish**	tur	2,062,037	23,926	86.18	0.934%
19	**Serbo-Croatian**	hbs	2,030,039	10,901	186.23	0.425%
20	**Ukrainian**	ukr	1,839,988	10,028	183.49	0.391%

Table S2 Statistics for the twenty languages with the most edits in our Wikipedia dataset. The full table is available in the SOM.

S1.3 Book translations

The Index Translationum is an international bibliography of book translations maintained by UNESCO[13]. The online database contains information on books translated and published in print in about 150 countries since 1979. Some countries are missing data for certain years, such as the United Kingdom in the years 1995-2000 and 2009-2011[39].

We 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, we converted the language codes listed in the Index Translationum to standard three-letter ISO639-3 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 ISO639-3 codes (together accounting for about a thousand translations total), and 5 administrative codes (*mis, mul, und, zxx,* and *not supplied*; see ISO639-3 documentation[15]). The remaining languages were merged into macrolanguages (see **Section S2.1**).

Table S3 shows statistics for the languages with the most translations in our dataset. The final dataset contains 2,231,920 translations in 1,019 languages. We used this dataset to generate the book translations GLN shown in **Figure 1** of the main section.

#	Language	Code	Translations from	Translations to	Total translations
1	**English**	eng	1,225,237	146,294	1,371,531
2	**German**	deu	201,718	292,124	493,842
3	**French**	fra	216,624	238,463	455,087
4	**Spanish**	spa	52,955	228,910	281,865
5	**Russian**	rus	101,395	82,772	184,167
6	**Japanese**	jpn	26,921	130,893	157,814
7	**Dutch**	nld	18,978	111,371	130,349
8	**Italian**	ita	66,453	59,830	126,283
9	**Swedish**	swe	39,192	71,688	110,880
10	**Polish**	pol	14,104	76,720	90,824
11	**Portuguese**	por	11,390	74,721	86,111
12	**Danish**	dan	21,239	64,799	86,038
13	**Czech**	ces	17,202	64,442	81,644
14	**Chinese**	zho	13,337	62,650	75,987
15	**Hungarian**	hun	11,256	54,989	66,245
16	**Norwegian**	nor	14,530	45,923	60,453
17	**Serbo-Croatian**	hbs	12,743	45,036	57,779
18	**Finnish**	fin	8,296	46,271	54,567
19	**Modern Greek (1453-)**	ell	4,862	27,422	32,284
20	**Bulgarian**	bul	3,667	25,742	29,409

Table S3 Statistics for the twenty languages with the most translations (to and from) in our Index Translationum dataset. The full table is available in the SOM.

S2 Language notation and demographics

S2.1 Notation

Each of our three datasets uses a different system for identifying language names. For the sake of consistency, we 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[15]. 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 we do not consider their speakers as polyglots. Instead, we merged mutually intelligible languages to macrolanguages following the ISO 639-3 Macrolanguage Mappings[15]. For example, we merged 29 varieties of Arabic into one Arabic macrolanguage (*ara*), and Malaysian, Indonesian, and 34 other Bhasa languages into a Malay macrolanguage (*msa*).

Another reason for consolidating languages is that the language detector we 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, we added a couple of merges that are not in the ISO 639-3 macrolanguage mappings: we 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*. Our full conversion table is available on the SOM page*.*

Languages belong to language families[40]. We mapped languages to language families using the hierarchy in Ethnologue[41] complemented by information from articles from the English Wikipedia about the respective languages. We used the standard language family names and identifiers as defined by ISO 639-5[42].

S2.2 Population

We use language speaker estimates from the June 14, 2012 version of Wikipedia Statistics page[16]. These estimates include all speakers of a language, native and non-native alike. We converted language names to ISO 639-3 identifiers and merged them into macrolanguages as explained in Section S2.1.

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 S4** we 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. Refer to **Table S5** for number of speakers for the languages in our 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) Twi (tw)	19 million 15 million
Arabic	ara	530 million	Arabic (ar) Egyptian Arabic (arz)	530 million 76 million
Malay	msa	300 million	Malay (ms) Indonesian (id)	300 million 250 million
Serbo-Croatian	hbs	23 million	Serbo-Croatian (sh) Serbian (sr) Croatian (hr) Bosnian (bs)	23 million 23 million 6 million 3 million
Norwegian	nor	5 million	Norwegian (no) Nynorsk (nn)	5 million 5 million
Komi	kom	293,000	Komi (kv) Komi-Perniak (koi)	293,000 94,000

Table S4 Macrolanguages for which the estimated number of speakers is not the sum of the estimates for the individual languages that constitute them.

	Language	Code	Speakers (millions)	GDP per capita ($)
1	**Afrikaans**	afr	13	10,373
2	**Albanian**	sqi	16	9,182
3	**Arabic**	ara	530	8,720
4	**Armenian**	hye	6	5,598
5	**Azerbaijani**	aze	27	11,902
6	**Bashkir**	bak	2	
7	**Basque**	eus	1	30,626
8	**Belarusian**	bel	6	15,028
9	**Bengali**	ben	230	2,457
10	**Bulgarian**	bul	12	13,488
11	**Catalan**	cat	9	30,626
12	**Chinese**	zho	1575	9,207
13	**Czech**	ces	12	27,062
14	**Danish**	dan	6	37,152
15	**Dutch**	nld	27	40,518
16	**English**	eng	1500	32,953
17	**Esperanto**	epo	1	
18	**Estonian**	est	1.07	20,380
19	**Filipino**	fil	90	4,073
20	**Finnish**	fin	6	36,236
21	**French**	fra	200	15,103
22	**French (Old)**	fro		
23	**Galician**	glg	4	30,626
24	**Georgian**	kat	4	5,491
25	**German**	deu	185	38,268
26	**German (Middle High)**	gmh		
27	**Greek (Ancient)**	grc		
28	**Greek (Modern)**	ell	15	26,693
29	**Haitian**	hat	12	1,235
30	**Hebrew**	heb	10	30,975
31	**Hindi**	hin	550	3,696
32	**Hungarian**	hun	15	18,672
33	**Icelandic**	isl	0.32	38,061
34	**Italian**	ita	70	30,623
35	**Japanese**	jpn	132	34,740
36	**Kara-Kalpak**	kaa	0.41	
37	**Kazakh**	kaz	12	13,001
38	**Kirghiz**	kir	5	2,372
39	**Korean**	kor	78	21,723
40	**Latin**	lat	0.01	
41	**Latvian**	lav	2.15	15,662
42	**Lithuanian**	lit	4	18,856
43	**Macedonian**	mkd	3	10,367
44	**Malay**	msa	300	6,023
45	**Malayalam**	mal	37	3,694
46	**Maltese**	mlt	0.37	25,428
47	**Maori**	mri	0.157	27,668
48	**Marathi**	mar	90	3,694
49	**Moldavian**	mol	3.5	
50	**Mongolian**	mon	5	4,744
51	**Norwegian**	nor	5	53,471
52	**Occitan**	oci	2	
53	**Persian**	fas	107	9,826
54	**Polish**	pol	43	20,326
55	**Portuguese**	por	290	11,853
56	**Romanian**	ron	28	11,354
57	**Russian**	rus	278	15,487
58	**Sanskrit**	san	0.05	
59	**Serbo-Croatian**	hbs	23	12,908
60	**Sinhala**	sin	19	5,674
61	**Slovak**	slk	7	23,432
62	**Slovenian**	slv	2	28,642
63	**Spanish**	spa	500	16,777
64	**Swahili**	swa	50	1,415
65	**Swedish**	swe	10	40,265
66	**Tajik**	tgk	4	2,238
67	**Tamil**	tam	66	3,923
68	**Tatar**	tat	8	
69	**Thai**	tha	73	9,396
70	**Tibetan**	bod	7	
71	**Turkish**	tur	70	14,623
72	**Turkmen**	tuk	9	5,816
73	**Uighur**	uig	10	
74	**Ukrainian**	ukr	45	7,242
75	**Urdu**	urd	60	3,511
76	**Uzbek**	uzb	24	3,182
77	**Vietnamese**	vie	80	3,447
78	**Welsh**	cym	0.75	
79	**Yiddish**	yid	3	

Table S5 Population and GDP per capita for the languages used in the GLNs. Blank cells indicate dead languages or insufficient data.

S2.3 Language GDP

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 Davis[1]. Given a country c, let G_c be the GDP per capita (based on purchasing-power-parity) of that country (2011 values; retrieved from the IMF[43] with a few additions from the CIA World Factbook[44]). Also, given a language l, let N_{lc} be the number of native speakers of l in country c, obtained from Ethnologue[41] and The World Factbook[44]. We calculated N_{lc} using the language demographics listed in **Table S8**. Thus, G_l, the GDP per capita for l is

$$G_l = \frac{\sum_c (G_c N_{lc})}{\sum_c N_{lc}}$$

Refer to **Table S5** for GDP per capita for the languages in our GLNs. These values are an approximation because the economic activity of a country is not distributed evenly by language. Moreover, a person may contribute in a language different than his or her native language: for example, many use English to communicate at their workplace although English is not their native language. Tables of GDP per capita and population by country and language are available on the SOM page.

S3 Additional calculations

In this section we briefly document two calculations used in the main text of the paper. First, we note that for all figures we use the number of multilingual speakers, or expressions, from a language. We estimate the number of multilingual speakers or expression from a language (N_i) as:

$$N_i = \sum_j M_{ij}$$

Also, we note that we estimate the eigenvector centrality of a language by using:

$$\lambda v_i = \sum_j M_{ij} v_j$$

and finding the eigenvector *v*, associated with the largest eigenvalue. Since the eigenvector associated with the largest eigenvalue could be positive or negative, we take the absolute value of the elements of this eigenvector as our measure of a language's eigenvector centrality.

S4 Language centrality: alternatives and robustness

S4.1 Eigenvector centrality vs. betweenness centrality

In this section we compare two measures of centrality the eigenvector centrality metric used in the main text with betweenness centrality. The *betweenness centrality* of a node is the number of shortest paths from all nodes to all others that pass through that node.[45] This centrality value focuses on quantity rather than quality: all shortest paths that go through a node contribute equally to its betweenness score, regardless of the characteristics of the source and target nodes (e.g., the number of their neighbors or their identity). The *eigenvector centrality* of a node is the sum of its summed connections to others, weighted by their centralities[26]. Eigenvector centrality thus takes into account the quality of a node's connections, by rewarding a node for being connected to "important" nodes. Each node is assigned a relative score based on its connections, and a connection to a high-scoring node contributes more to the eigenvector centrality score of the node being scored than a connection to a low-scoring node.

Figure S3 and **Table S6** show the correlation of eigenvector centrality and betweenness centrality for all languages and datasets. The correlation between the two centrality measures is R^2=0.25 for Twitter, R^2=0.62 for Wikipedia, and R^2=0.39 for book translations.

The deviations between these two centrality measures are quite informative. For instance, according to betweenness centrality the most central language in the book translations GLN is Russian. **Figure 1** in the main text shows why: Russian is the portal to a large number of languages that would otherwise be disconnected from the rest of the network (such as Tatar, Armenian and Kirghiz). All paths to these languages pass through Russian, contributing to Russian's high betweenness score. The same is not true for English, the language with the second-highest betweenness. While English is also highly connected, it is connected to many languages that are connected to others, and is thus located in a part of the network where there are alternative paths that reduce the betweenness of English. At the same time, the fact that English is connected to languages that are connected to others increases its eigenvector centrality.

We chose eigenvector centrality over betweenness, as the former is more suitable for identifying global languages according to our definition: a global language is a language that are connected to other hub languages (such as English in the example from the book translations network above), not a language that serve as the only gateway to many peripheral languages (such as Russian in the above example).

We also had a practical reason for preferring eigenvector centrality over betweenness centrality: the latter is a measure that is unable to differentiate among more peripheral languages, since most languages get a betweenness score of zero (see **Figure S3** below). Eigenvector centrality, on the other hand, can help us differentiate between the positions of languages in the GLN at all levels of centrality, not only among the most central languages.

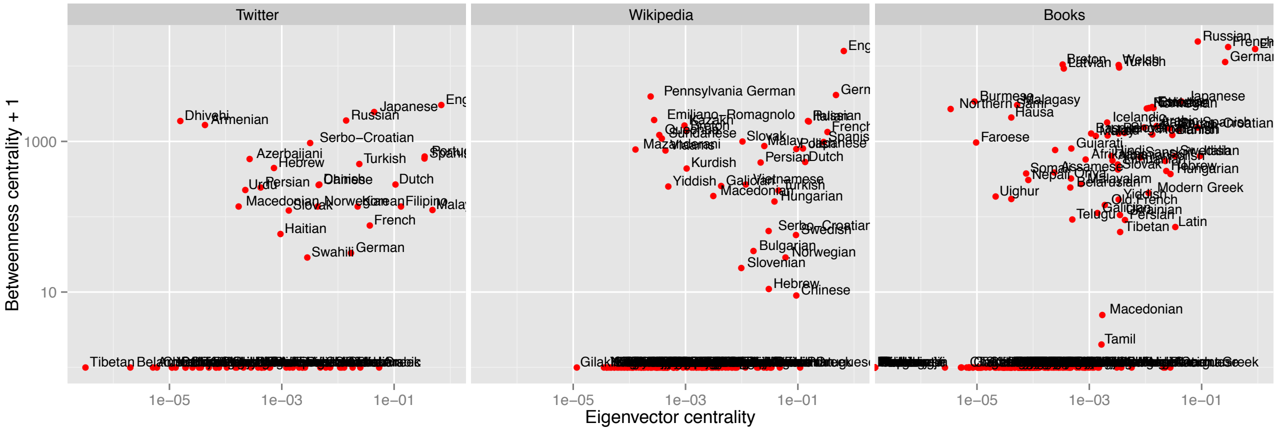


Figure S3 Comparison between eigenvector centrality and betweenness centrality, calculated as the total number of paths going through a node, for **A** The Twitter GLN **B** The Wikipedia GLN **C** The book translations GLN.

		Eigenvector centrality			Betweenness centrality		
Language	Code	Twitter	Wikipedia	Books	Twitter	Wikipedia	Books
Abaza	abq			5.22E-06			0
Abkhazian	abk			6.92E-05			0
Adyghe	ady			4.64E-05			0
Afrikaans	afr	0.01171827	0.00223104	0.00086109	0	0	579
Akkadian	akk			0.00021226			0
Albanian	sqi	0.00259932	0.00688492	0.00260489	0	0	559
Amharic	amh	5.10E-06	0.0002177	0	0	0	0
Ancient Greek (to 1453)	grc			0.02250273			0
Anglo-Norman	xno			5.48E-05			0
Arabic	ara	0.05422551	0.03207688	0.01588606	0	0	1598
Aragonese	arg		0.00021357	0		0	0
Armenian	hye	4.32E-05	0.00172884	0.00149675	1656	0	0
Assamese	asm		0.00024102	0.00023965		0	389
Asturian	ast		0.00036611	3.19E-05		0	0
Avaric	ava			8.82E-05			0
Azerbaijani	aze	0.00026921	0.00301522	0.00028409	586	0	0
Azhe	yiz			0			0
Bambara	bam			5.66E-05			0
Bashkir	bak			0.00018743			0
Basque	eus	0.00061809	0.0031492	0.00108321	0	0	1273
Bavarian	bar		0.00193201			0	
Belarusian	bel	2.02E-06	0.00127629	0.00046173	0	0	244
Bengali	ben	5.05E-05	0.0039574	0.00334541	0	0	1280
Biali	beh			0			0
Bikol	bik		0.00011662			0	
Breton	bre		0.00103388	0.00033737		1368	10475
Buamu	box			0			0
Buduma	bdm			2.94E-05			0
Bulgarian	bul	0.00028475	0.01627752	0.00371912	0	34	0
Buriat	bua			7.25E-05			0
Burmese	mya	1.30E-05	0.00041984	9.06E-06	0	0	3379
Caribbean Hindustani	hns			6.64E-06			0
Caribbean Javanese	jvn			2.71E-06			0
Catalan	cat	0.010975	0.02784782	0.00949586	0	0	1538
Cebuano	ceb		0.00034209			0	
Central Atlas Tamazight	tzm			9.74E-05			0
Central Khmer	khm	1.11E-05	0.00043539	8.22E-05	0	0	0
Cerma	cme			0			0
Chechen	che			4.05E-05			0
Cherokee	chr	6.02E-06			0		
Chinese	zho	0.00453649	0.09375027	0.01396375	264	8	2793
Chipewyan	chp			0.00012325			0
Chukot	ckt			1.24E-05			0
Church Slavic	chu			5.34E-06			0
Chuvash	chv		6.44E-05	0.00011037		0	0
Classical huatl	nci			0.00019172			0
Coptic	cop			0.00040399			0
Cornish	cor		0.00017882	5.48E-05		0	0
Corsican	cos		8.37E-05	9.96E-05		0	0
Cree	cre			0.00046563			0
Crimean Tatar	crh			1.18E-05			0
Czech	ces	0.01293374	0.03311951	0.02775867	0	0	0
Dakota	dak			4.11E-05			172
Danish	dan	0.00467135	0.03408103	0.03020216	268	0	1219
Dargwa	dar			3.98E-05			0
Dhivehi	div	1.56E-05			1858		
Djeebba	djj			0.00012325			0
Dogosé	dos			0			0
Dolgan	dlg			7.18E-06			0
Dutch	nld	0.10582998	0.13486947	0.03955701	268	530	1414
Eastern Maroon Creole	djk			7.24E-06			0
Egyptian (Ancient)	egy			0.00036975			0
Emiliano-Romagnolo	eml		0.00027622			1940	
English	eng	0.69329476	0.65929841	0.89803531	3051	15782	17006
Erzya	myv			3.79E-05			0
Esperanto	epo		0.00827705	0		0	0
Estonian	est	0.00044706	0.00663579	0.01322228	0	0	2879
Even	eve			0			0
Evenki	evn			1.18E-05			0

		Eigenvector centrality			Betweenness centrality		
Language	**Code**	**Twitter**	**Wikipedia**	**Books**	**Twitter**	**Wikipedia**	**Books**
Faroese	fao		0.00041207	9.67E-06		0	970
Fiji Hindi	hif		0.00024102			0	
Filipino (macrolanguage)	fil	0.13316405	0.00375525	0.00045877	136	0	0
Finnish	fin	0.00138618	0.04780708	0.01160693	0	0	2762
French	fra	0.03696668	0.33783444	0.29695532	76	1338	17924
Friulian	fur		7.41E-05			0	
Fulah	ful			3.85E-05			0
Fuliiru	flr			0			0
Gagauz	gag			7.18E-06			0
Galician	glg	0.00486003	0.00433051	0.0014028	0	254	111
Gascon	gsc			2.49E-05			0
Geez	gez			5.48E-05			0
Georgian	kat	0.00010531	0.00239351	0.00043495	0	0	0
German	deu	0.01711333	0.4787209	0.26334749	32	4144	11331
Gilaki	glk		1.16E-05			0	
Guadeloupean Creole French	gcf			0.00024756			772
Guarani	grn		9.15E-05			0	
Guianese Creole French	gcr			3.62E-05			0
Gujarati	guj	1.93E-05	0.00090966	0.00048148	0	0	801
Haitian	hat	0.00095278		7.53E-05	58		0
Hani	hni			0			0
Hausa	hau			4.11E-05			2070
Hawaiian	haw			0.0001164			0
Hebrew	heb	0.00072731	0.03049572	0.02361634	440	10	403
Hindi	hin	0.00043965	0.00575338	0.00247332	0	0	641
Hittite	hit			4.11E-05			0
Hmong	hmn			3.42E-05			0
Hopi	hop			0.00010271			0
Hungarian	hun	0.00121054	0.03845013	0.02802628	0	158	370
Icelandic	isl	0.00015844	0.00390985	0.00208855	0	0	1778
Ido	ido		0.00032654			0	
Ifè	ife			0			0
Ingush	inh			2.94E-05			0
Interlingua (Intertiol Auxiliary Language Association)	ina		0.00044317			0	
Inuktitut	iku			0.00013131			0
Iri Sami	smn			0			0
Irish	gle	0.00015056	0.00195926	0.00082199	0	0	0
Italian	ita	0.02517954	0.15703838	0.09374308	0	1828	637
Japanese	jpn	0.04418507	0.12399229	0.04398496	2465	802	3369
Javanese	jav		0.00061884			0	
Jola-Fonyi	dyo			0			0
Kabardian	kbd			6.66E-05			0
Kabyle	kab		0.00011155	8.60E-05		0	0
Kachin	kac			0			0
Kalaallisut	kal		0.00010885	1.45E-05		0	0
Kalmyk	xal			8.88E-05			0
Kanda	kan	3.62E-05	0.00195926	0.00066418	0	0	0
Kara-Kalpak	kaa			6.20E-05			0
Karachay-Balkar	krc			7.71E-05			0
Karang	kzr			0			0
Karelian	krl			1.11E-05			0
Kasem	xsm			0			0
Kashmiri	kas		4.67E-05	4.11E-05		0	0
Kashubian	csb		3.54E-05			0	
Kazakh	kaz		0.000961	0.00043691		1624	0
Khakas	kjh			1.70E-05			0
Khanty	kca			1.05E-05			0
Kikuyu	kik			6.85E-05			0
Kinyarwanda	kin		5.44E-05			0	
Kirghiz	kir		4.83E-05	0.00023903		0	0
Kölsch	ksh		0.0003105			0	
Komi	kom			5.75E-05			0
Korean	kor	0.02250541	0.02475757	0.00309892	136	0	0
Kriol	rop			4.11E-05			0
Kumyk	kum			2.42E-05			0
Kurdish	kur		0.00105145	0		440	0
Ladino	lad			8.22E-05			0
Lahu	lhu			0			0

		Eigenvector centrality			Betweenness centrality		
Language	**Code**	**Twitter**	**Wikipedia**	**Books**	**Twitter**	**Wikipedia**	**Books**
Lak	lbe			2.22E-05			0
Lama (Togo)	las			0			0
Lao	lao	2.59E-05	0.00016327	6.85E-05	0	0	0
Latin	lat		0.00785178	0.03404642		0	72
Latvian	lav	0.00234029	0.00346512	0.00035201	0	0	9378
Lezghian	lez			2.09E-05			0
Ligurian	lij		5.37E-05			0	
Limburgan	lim		0.0004847			0	
Lingala	lin			2.94E-05			0
Lisu	lis			0			0
Lithuanian	lit	0.0002094	0.0075504	0.00322933	0	0	491
Lombard	lmo		0.00047653			0	
Low German	nds		0.0013855			0	
Lule Sami	smj			0			0
Luxembourgish	ltz		0.00235924			0	
Macedo-Romanian	rup		0.00014772			0	
Macedonian	mkd	0.00017164	0.0031274	0.00171642	136	188	4
Malagasy	mlg			5.21E-05			3039
Malay (macrolanguage)	msa	0.48559568	0.02524643	0.00131467	122	864	1183
Malayalam	mal	6.72E-05	0.0051936	0.00071301	0	0	273
Maltese	mlt	0.00060428	0.00060644	7.53E-05	0	0	0
Maori	mri			0.00034236			0
Marathi	mar		0.00204478	0.00086517		0	0
Mari (Russia)	chm			6.34E-05			0
Mazanderani	mzn		0.00012852			780	
Middle Dutch (ca. 1050-1350)	dum			0.00011357			0
Middle English (1100-1500)	enm			0.00102024			0
Middle French (ca. 1400-1600)	frm			0.00075791			0
Middle High German (ca. 1050-1500)	gmh			0.00120464			0
Middle Irish (900-1200)	mga			6.16E-05			0
Modern Greek (1453-)	ell	0.0061217	0.01877043	0.01143029	0	0	205
Mofu-Gudur	mif			0			0
Moksha	mdf			5.75E-05			0
Moldavian	mol		0.00011662	0.00570555		0	0
Mongolian	mon		0.00069974	4.25E-05		0	0
Muyang	muy			0			0
Nanai	gld			1.11E-05			0
Navajo	nav		0.00010107	5.48E-05		0	0
Neapolitan	nap		0.00039139			0	
Nenets	yrk			2.16E-05			0
Nepali (macrolanguage)	nep		0.00060644	8.22E-05		0	307
Nepali macrolanguage	nep		0.00060644	8.22E-05		0	307
Newari	new			0			0
Ngangam	gng			0			0
Ngiemboon	nnh			0			0
Nogai	nog			1.18E-05			0
Northern Sami	sme			3.41E-06			2680
Northern Yukaghir	ykg			1.05E-05			0
Norwegian	nor	0.00436981	0.06036022	0.0106394	136	28	2755
Occitan (post 1500)	oci		0.00029083	0.00095096		0	0
Official Aramaic (700-300 BCE)	arc			0.00113664			0
Ojibwa	oji			6.16E-05			0
Old English (ca. 450-1100)	ang		0.00039652	0.00060256		0	0
Old French (842-ca. 1400)	fro			0.00189681			142
Old High German (ca. 750-1050)	goh			3.01E-05			0
Old Irish (to 900)	sga			6.16E-05			0
Old Japanese	ojp			8.49E-06			0
Old Norse	non			0.00028642			0
Old Provençal (to 1500)	pro			0.00018114			0
Old Russian	orv			6.73E-05			0
Oriya (macrolanguage)	ori		0.00025657	0.00047401		0	323
Oriya macrolanguage	ori		0.00025657	0.00047401		0	323
Ossetian	oss			8.88E-05			0
Ottoman Turkish (1500-1928)	ota			4.79E-05			0
Pahlavi	pal			4.11E-05			0
Pali	pli			0.00080113			0
Pampanga	pam		9.33E-05			0	
Panjabi	pan	7.66E-05	0.00031099	0.00085128	0	0	0
Papiamento	pap		0.00019437			0	

		Eigenvector centrality			Betweenness centrality		
Language	**Code**	**Twitter**	**Wikipedia**	**Books**	**Twitter**	**Wikipedia**	**Books**
Pennsylvania German	pdc		0.00024018			3960	
Persian	fas	0.00042093	0.02186386	0.0043465	242	526	90
Picard	pcd			4.08E-05			0
Polish	pol	0.00234695	0.0938191	0.02271051	0	796	546
Portuguese	por	0.35208674	0.10986822	0.02105229	628	0	0
Pushto	pus		0.00013217	0		0	0
Quechua	que		0.00034272	0		1218	0
Rajasthani	raj			5.48E-05			0
Réunion Creole French	rcf			7.25E-05			0
rom	nrm		3.98E-05			0	
Romanian	ron	0.00157288	0.02808689	0.01328133	0	0	1241
Romansh	roh		0.00030441	0.00010843		0	0
Romany	rom			0			0
Russian	rus	0.01401232	0.1516584	0.08565274	1902	1864	21213
Sanskrit	san		0.00062976	0.00815953		0	610
Saramaccan	srm			7.54E-06			0
Sardinian	srd		7.96E-05			0	
Scots	sco		0.00107293	8.90E-05		0	0
Scottish Gaelic	gla		0.00046649	0.0002602		0	0
Serbo-Croatian	hbs	0.003212	0.03030069	0.02696108	950	64	1482
Sichuan Yi	iii			0			0
Sicilian	scn		0.00071389			0	
Silesian	szl		4.32E-05			0	
Sindhi	snd			9.59E-05			0
Sinhala	sin	6.86E-05	0.00155497	0.00015749	0	0	0
Slovak	slk	0.00134281	0.01036716	0.00325322	120	1004	424
Slovenian	slv	0.0004879	0.00987944	0.00428489	0	20	1302
Somali	som		0.00046649	7.53E-05		0	377
Southern Altai	alt			3.14E-05			0
Southern Sami	sma			0			0
Spanish	spa	0.34811446	0.28746349	0.08539987	594	982	1531
Sran Tongo	srn			1.03E-05			0
Sumerian	sux			0.00016433			0
Sundanese	sun		0.00037938			1088	
Swahili (macrolanguage)	swa	0.00285522	0.00069974	8.90E-05	28	0	0
Swati	ssw			0			0
Swedish	swe	0.00728292	0.0927484	0.03363697	0	56	640
Swiss German	gsw			3.41E-05			0
Syriac	syr			0.00047246			0
Tabassaran	tab			8.49E-06			0
Tai Hongjin	tiz			0			0
Tajik	tgk			0.00018351			0
Tamashek	tmh			4.76E-05			0
Tamil	tam	0.00043373	0.00445498	0.00165045	0	0	1
Tatar	tat		6.08E-05	0.00022792		0	0
Tavringer Romani	rmu			0			0
Telugu	tel	3.36E-05	0.00285337	0.00050067	0	0	91
Tembo (Kitembo)	tbt			2.72E-05			0
Tepo Krumen	ted			0			0
Thai	tha	0.0183352	0.01166767	0.00030757	0	0	0
Thayore	thd			0			0
Tibetan	bod	3.20E-07		0.00356449	0		62
Tok Pisin	tpi		0.00016327			0	
Tokelau	tkl			0.00010271			0
Turkish	tur	0.02408795	0.0451629	0.00340584	504	222	9669
Turkmen	tuk			0.00016915			0
Tuvinian	tyv			3.00E-05			0
Udmurt	udm			5.75E-05			0
Uighur	uig			2.16E-05			185
Ukrainian	ukr	1.95E-05	0.01974593	0.00352934	0	0	105
Urdu	urd	0.00022531	0.00157829	0.00212528	226	0	1203
Uzbek	uzb			0.00033829			0
Venetian	vec		0.00018149			0	
Vietmese	vie	0.00204518	0.01183458	0.00142575	0	268	0
Vlaams	vls		0.00043929			762	
Wa	wbm			0			0
Walloon	wln			0			0
Warlpiri	wbp			0.00010271			0
Welsh	cym	0.00038081	0.00212253	0.00334156	0	0	10390

		Eigenvector centrality			Betweenness centrality		
Language	**Code**	**Twitter**	**Wikipedia**	**Books**	**Twitter**	**Wikipedia**	**Books**
Western Frisian	fri			3.08E-05			0
Western Frisian	fry		0.00016223			0	
Wolof	wol			0			0
Xhosa	xho			0			0
xi	nbf			0			0
Yakut	sah		3.94E-05	8.88E-05		0	0
Yiddish	yid	1.11E-05	0.00049499	0.00334887	0	252	169
Yoruba	yor			9.59E-05			0
Zhuang	zha			0			0
Zulu	zul		7.00E-05	7.53E-05		0	0

Table S6 Eigenvector and betweenness centrality by language in each of the three GLNs (rounded to the nearest hundredth).

S5 Famous people per language

We measure the cultural impact of a language by the number of its speakers who made a long-lasting cultural impression on the world. We focus on these *famous* people, rather than on ideas or other forms of cultural expression, because people names are easier to identify and match across languages.

We use two separate methods to decide whether a person is famous. The first is having Wikipedia articles in at least 26 language editions, and the second is being included in the *Human Accomplishment* list[27], a list of nearly 4,000 influential people in the arts and sciences, from 800 BCE to 1950. As neither dataset contains information about the language used by the famous people it lists, we start this section by describing how we associated famous people with languages. Then, we dedicate a subsection to each dataset, in which we describe how the dataset was retrieved and prepared for use.

S5.1 Associating a famous person 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, we determined a person's language affiliation using the current language demographics for his or her country of birth. Each famous person in our datasets equals one point, which is distributed across the languages spoken in his or her native country according to their population[41,44]. For example, Italian inventor Guglielmo Marconi counts as one point for Italian. Former Canadian Prime Minister Pierre Trudeau contributes 0.59 to English and 0.22 to French. We stress again that our 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 0.59 points for English and 0.22 points for French, regardless of their native language. Refer to **Tables S10a-b** for the language demographics of each country.

We determine a person's country of birth using present-day international borders. For example, we 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. In some cases, 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, 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, we also set 1950 as the latest year of birth for the Wikipedia dataset.

Despite some inaccuracies, using present-day countries provides a consistent mapping of people who lived over a period of several millennia to their contemporary countries. Moreover, using present-day countries allows us 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.

1	**Afghanistan**	Persian 50%, Pushto 35%, Uzbek 6%, Turkmen 5%
2	**Albania**	Albanian 95%, Greek (Modern) 3%
3	**Algeria**	Arabic 80%, French 20%
4	**Andorra**	Catalan 40%, Spanish 35%, Portuguese 15%, French 5.5%
5	**Angola**	Portuguese 70%
6	**Argentina**	Spanish 98%
7	**Armenia**	Armenian 97.7%, Russian 0.9%
8	**Aruba**	Papiamento 66.3%, Spanish 12.6%, English 7.7%, Dutch 5.8%
9	**Australia**	English 78.5%, Chinese 2.5%, Italian 1.6%, Greek (Modern) 1.3%, Arabic 1.2%, Vietnamese 1%
10	**Austria**	German 88.6%, Serbo-Croatian 3.8%, Turkish 2.3%
11	**Azerbaijan**	Azerbaijani 90.3%, Lezghian 2.2%, Russian 1.8%, Armenian 1.5%
12	**Bahamas, The**	English 100%
13	**Bahrain**	Arabic 100%
14	**Bangladesh**	Bengali 98%
15	**Barbados**	English 100%
16	**Belarus**	Russian 70.2%, Belarusian 23.4%
17	**Belgium**	Dutch 60%, French 40%
18	**Belize**	English 41%, Spanish 32%
19	**Benin**	French 40%, Fon 39%, Yoruba 12%
20	**Bermuda**	English 100%
21	**Bhutan**	Tshangla 28%, Dzongkha 24%, Nepali 22%
22	**Bolivia**	Spanish 60.7%, Quechua 21.2%, Aymara 14.6%
23	**Bosnia and Herzegovina**	Serbo-Croatian 100%
24	**Botswana**	Tswana 78.2%, Kalanga 7.9%, English 2.1%
25	**Brazil**	Portuguese 100%
26	**Brunei**	Malay 100%
27	**Bulgaria**	Bulgarian 76.8%, Turkish 8.2%, Romany 3.8%
28	**Burkina Faso**	French 100%
29	**Burma**	Burmese 100%
30	**Burundi**	French 50%, Rundi 50%
31	**Cambodia**	Central Khmer 95%
32	**Cameroon**	French 50%, English 50%
33	**Canada**	English 58.8%, French 21.6%
34	**Cape Verde**	Portuguese 100%
35	**Central African Republic**	Sango 80%, French 20%
36	**Chad**	Arabic 50%, French 50%
37	**Chile**	Spanish 100%
38	**China**	Chinese 100%
39	**Colombia**	Spanish 100%
40	**Congo, Democratic Republic of the**	French 33%, Swahili 20%, Lingala 20%
41	**Congo, Republic of the**	French 30%, Ibali Teke 17%, Lingala 13%
42	**Costa Rica**	Spanish 100%
43	**Cote d'Ivoire**	French 50%, Baoulé 14%
44	**Croatia**	Serbo-Croatian 100%
45	**Cuba**	Spanish 100%
46	**Cyprus**	Greek (Modern) 77%, Turkish 18%
47	**Czech Republic**	Czech 95.4%, Slovak 1.6%
48	**Denmark**	Danish 100%
49	**Djibouti**	Somali 38%, Arabic 20%, French 20%, Afar 13%
50	**Dominican Republic**	Spanish 100%
51	**Ecuador**	Spanish 100%
52	**Egypt**	Arabic 100%
53	**El Salvador**	Spanish 100%
54	**Equatorial Guinea**	Spanish 67.6%, French 20%
55	**Eritrea**	Tigrinya 55%, Tigre 16%
56	**Estonia**	Estonian 67.3%, Russian 29.7%
57	**Ethiopia**	Oromo 33.8%, Amharic 29.3%, Somali 6.2%, Tigre 5.9%, Sidamo 4%
58	**Faroe Islands**	Faroese 100%
59	**Fiji**	Fiji Hindi 45.3%, Fijian 39.3%
60	**Finland**	Finnish 91.2%, Swedish 5.5%
61	**France**	French 100%
62	**French Guiana**	French 100%
63	**Gabon**	French 75%, Fang 25%
64	**Gambia, The**	English 100%
65	**Georgia**	Georgian 71%, Russian 9%, Armenian 7%, Azerbaijani 6%
66	**Germany**	German 100%
67	**Ghana**	Akan 24.7%, English 21.3%, Ewe 12.7%, Abron 4.6%
68	**Gibraltar**	English 100%
69	**Greece**	Greek (Modern) 99%
70	**Greenland**	Danish 100%
71	**Grenada**	English 87%, French 2%
72	**Guadeloupe**	French 99%
73	**Guam**	English 38.3%, Chamorro 22.2%, Filipino 22.2%
74	**Guatemala**	Spanish 60%
75	**Guinea**	French 100%
76	**Guinea-Bissau**	Upper Guinea Crioulo 44%, Portuguese 14%
77	**Guyana**	English 50%
78	**Haiti**	Haitian 75%, French 25%
79	**Honduras**	Spanish 100%
80	**Hong Kong**	Chinese 95%, English 3.5%
81	**Hungary**	Hungarian 93.6%
82	**Iceland**	Icelandic 100%
83	**India**	Hindi 41%, Bengali 8.1%, Telugu 7.2%, Marathi 7%, Tamil 5.9%, Urdu 5%, Gujarati 4.5%, Kannada 3.7%, Oriya 3.2%, Malayalam 3.2%, Panjabi 2.8%
84	**Indonesia**	Malay 100%
85	**Iran**	Persian 53%, Azerbaijani 18%, Kurdish 10%, Luri 6%, Arabic 2%
86	**Iraq**	Arabic 80%, Kurdish 15%
87	**Ireland**	English 95%, Irish 2%
88	**Isle of Man**	English 100%
89	**Israel**	Hebrew 80%, Arabic 15%
90	**Italy**	Italian 100%
91	**Jamaica**	English 100%
92	**Japan**	Japanese 100%
93	**Jersey**	English 94.5%, Portuguese 4.6%
94	**Jordan**	Arabic 100%
95	**Kazakhstan**	Kazakh 63%, Russian 24%
96	**Kenya**	Swahili 80%, English 20%
97	**Kiribati**	Gilbertese 62.6%
98	**Korea, North**	Korean 100%
99	**Korea, South**	Korean 100%
100	**Kosovo**	Albanian 100%

Table S10a Language demographics by country. Values for each country add to 100% or less.

101	**Kuwait**	Arabic 100%	126	**Morocco**	Arabic 90%	151	**Russia**	Russian 100%	176	**Taiwan**	Chinese 100%
102	**Kyrgyzstan**	Kirghiz 64.7%, Uzbek 13.6%, Russian 12.5%	127	**Mozambique**	Makhuwa 25.3%, Portuguese 10.7%, Tsonga 10.3%, Sena 7.5%, Lomwe 7%, Chuwabu 5.1%	152	**Rwanda**	Kinyarwanda 98%	177	**Tajikistan**	Tajik 100%
103	**Laos**	Lao 100%	128	**Namibia**	Afrikaans 60%, German 32%, English 7%	153	**Saint Kitts and Nevis**	English 100%	178	**Tanzania**	Swahili 100%
104	**Latvia**	Latvian 58.2%, Russian 37.5%	129	**Nauru**	Nauru 100%	154	**Saint Lucia**	English 100%	179	**Thailand**	Thai 100%
105	**Lebanon**	Arabic 80%, French 20%	130	**Nepal**	Nepali 47.8%, Maithili 12.1%, Bhojpuri 7.4%	155	**Samoa**	Samoan 90%, English 10%	180	**Timor-Leste**	Tetum 36.6%, English 31.4%, Portuguese 23.5%
106	**Lesotho**	Southern Sotho 100%	131	**Netherlands**	Dutch 100%	156	**Saudi Arabia**	Arabic 100%	181	**Togo**	French 30%
107	**Liberia**	English 20%	132	**New Caledonia**	French 97%	157	**Senegal**	Wolof 70%, French 10%	182	**Tonga**	Tonga (Tonga Islands) 70%, English 30%
108	**Libya**	Arabic 95%	133	**New Zealand**	English 91.2%, Maori 3.9%, Samoan 2.1%, Chinese 2.1%, French 1.3%, Hindi 1.1%	158	**Serbia**	Serbo-Croatian 90.1%, Hungarian 3.8%, Romany 1.1%	183	**Trinidad and Tobago**	English 90%
109	**Lithuania**	Lithuanian 82%, Russian 8%, Polish 5.6%	134	**Nicaragua**	Spanish 97.5%	159	**Seychelles**	Seselwa Creole French 91%, English 4.9%	184	**Tunisia**	Arabic 100%
110	**Luxembourg**	Luxembourgish 77%, French 6%, German 4%, English 1%	135	**Niger**	Hausa 49.6%, Zarma 25.5%, Tamashek 8.4%, Fulah 8.3%, French 5%	160	**Sierra Leone**	Krio 90%	185	**Turkey**	Turkish 85.4%, Kurdish 12%, Arabic 1.2%
111	**Macedonia**	Macedonian 66.5%, Albanian 25.1%, Turkish 3.5%, Romany 1.9%, Serbo-Croatian 1.2%	136	**Nigeria**	English 30%	161	**Singapore**	Chinese 58.8%, English 23%, Malay 14.1%, Tamil 3.2%	186	**Turkmen-istan**	Turkmen 72%, Russian 12%, Uzbek 9%
112	**Madagascar**	French 70%, Malagasy 30%	137	**Norway**	Norwegian 100%	162	**Slovakia**	Slovak 83.9%, Hungarian 10.7%, Romany 1.8%, Ukrainian 1%	187	**Uganda**	Ganda 14%, English 8%
113	**Malawi**	Nyanja 70%, Yao 10.1%, Tumbuka 9.5%	138	**Oman**	Arabic 100%	163	**Slovenia**	Slovenian 91.1%, Serbo-Croatian 4.5%	188	**Ukraine**	Ukrainian 67%, Russian 24%
114	**Malaysia**	Malay 100%	139	**Pakistan**	Panjabi 48%, Sindhi 12%, Lahnda 10%, Urdu 8%, Pushto 8%	164	**Solomon Islands**	English 2%	189	**United Arab Emirates**	Arabic 100%
115	**Maldives**	Dhivehi 100%	140	**Palestinian Authority**	Arabic 100%	165	**Somalia**	Somali 80%, Arabic 20%	190	**United Kingdom**	English 100%
116	**Mali**	Bambara 46.3%, French 10%, Fulah 9.4%, Soninke 6.4%	141	**Panama**	Spanish 100%	166	**South Africa**	Zulu 23.82%, Xhosa 17.64%, Afrikaans 13.35%, Pedi 9.39%, Tswana 8.2%, English 8.2%, Southern Sotho 7.93%	191	**United States**	English 82.1%, Spanish 10.7%
117	**Malta**	Maltese 90.2%, English 6%	142	**Papua New Guinea**	English 2%, Tok Pisin 1.8%	167	**South Sudan**	Arabic 50%	192	**Uruguay**	Spanish 100%
118	**Martinique**	French 100%	143	**Paraguay**	Guarani 50%, Spanish 50%	168	**Spain**	Spanish 74%, Catalan 17%, Galician 7%, Basque 2%	193	**Uzbekistan**	Uzbek 74.3%, Russian 14.2%, Tajik 4.4%
119	**Mauritania**	Arabic 100%	144	**Peru**	Spanish 84.1%, Quechua 13%, Aymara 1.7%	169	**Sri Lanka**	Sinhala 74%, Tamil 18%	194	**Vanuatu**	Bislama 23.1%, English 1.9%, French 1.4%
120	**Mauritius**	Bhojpuri 12.1%, French 3.4%, English 1%	145	**Philippines**	Filipino 100%	170	**Sudan**	Arabic 100%	195	**Venezuela**	Spanish 100%
121	**Mexico**	Spanish 98.5%	146	**Poland**	Polish 97.8%	171	**Suriname**	Dutch 60%	196	**Vietnam**	Vietnamese 100%
122	**Moldova**	Romanian 76.5%, Russian 11.2%, Ukrainian 4.4%, Gagauz 4%, Bulgarian 1.6%	147	**Portugal**	Portuguese 100%	172	**Swaziland**	Swati 98%	197	**Virgin Islands**	English 74.7%, Spanish 16.8%, French 6.6%
123	**Monaco**	French 100%	148	**Puerto Rico**	Spanish 90%, English 10%	173	**Sweden**	Swedish 100%	198	**Yemen**	Arabic 100%
124	**Mongolia**	Mongolian 90%	149	**Qatar**	Arabic 100%	174	**Switzerland**	German 63.7%, French 20.4%, Italian 6.5%, Serbo-Croatian 1.5%, Albanian 1.3%, Portuguese 1.2%, Spanish 1.1%, English 1%	199	**Zambia**	Bemba 30.1%, English 16%, Nyanja 10.7%, Tonga (Zambia) 10.6%, Lozi 5.7%
125	**Montenegro**	Serbo-Croatian 91.1%, Albanian 5.3%	150	**Romania**	Romanian 91%, Hungarian 6.7%, Romany 1.1%	175	**Syria**	Arabic 100%	200	**Zimbabwe**	Shona 70%, North Ndebele 20%, English 2.5%

Table S10b Language demographics by country. Values for each country add to 100% or less.

S5.2 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[46,47]. 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.

We compiled our *Wikipedia* dataset of famous people as follows. We 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. We used a data dump from November 4, 2012; the latest version of the table is available from Freebase[48]. 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 we retrieved through the Freebase API[49]. The *pageid* and the Wikipedia API[50] were used to find the number of language editions in which a person had an article. Then, the pageid, Wikipedia article name, and number of languages of each article were added to the table retrieved from Freebase.

We matched 991,684 people with the English Wikipedia, from which we selected 216,280 people with a defined date of birth, place of birth and gender. We then restricted this list to include only the 11,340 people who had articles in at least 26 Wikipedia language editions and a defined date of birth, place of birth and gender. After examining biographical articles in all Wikipedia language editions, we found that there is no biography that appears in at least 26 languages or more that does not have an English version. Thus, by compiling biographies from the English Wikipedia we capture the famous people in any other Wikipedia language. The 26-language threshold generated a group that is exclusive enough while still containing enough data points. For comparison, a 20-language threshold would give us 13,334 articles, and a 30-language threshold would give us 6,336 articles.

Next, we converted dates to a standard four-digit year format. While doing so, we 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. We then used the Google Geocoding API[51] 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. After dropping records with an ambiguous place of birth we remained with 10,773 people—to which we refer henceforth as the *Wikipedia 26* dataset. Finally, we converted countries to languages as described in Section 4.1 above. To increase the accuracy of the conversion, we selected from the Wikipedia 26 dataset only the 4,886 people who were born after 1800 and before 1950.

Tables S11 and S12 show the number of famous people for each country and language, respectively.

	Country	People (all years)	People (1800-1950)
1	Afghanistan	21	10
2	Albania	15	7
3	Algeria	17	11
4	Andorra	1	
5	Angola	5	4
6	Antigua and Barbuda	1	1
7	Argentina	102	33
8	Armenia	12	4
9	Aruba	1	1
10	Australia	95	28
11	Austria	139	91
12	Azerbaijan	15	6
13	Bahrain	1	1
14	Bangladesh	8	7
15	Barbados	1	
16	Belarus	22	10
17	Belgium	103	40
18	Benin	3	1
19	Bermuda	1	
20	Bhutan	4	1
21	Bolivia	3	1
22	Bosnia and Herzegovina	26	8
23	Botswana	4	3
24	Brazil	137	53
25	Brunei	1	1
26	Bulgaria	29	8
27	Burkina Faso	2	1
28	Burma	7	7
29	Burundi	1	
30	Cambodia	5	2
31	Cameroon	11	2
32	Canada	106	46
33	Cape Verde	4	1
34	Central African Republic	1	1
35	Chad	2	
36	Chile	27	13
37	China	94	37
38	Colombia	17	3
39	Congo, Democratic Republic of the	7	3
40	Congo, Republic of	2	1
41	Costa Rica	3	1
42	Cote d'Ivoire	15	3
43	Croatia	56	10
44	Cuba	13	9
45	Cyprus	9	5
46	Czech Republic	105	53
47	Denmark	99	39
48	Djibouti	1	
49	Dominican Republic	2	1
50	Ecuador	4	1
51	Egypt	68	24
52	El Salvador	3	1
53	Equatorial Guinea	1	1
54	Eritrea	1	1
55	Estonia	15	9
56	Ethiopia	10	6
57	Faroe Islands	1	1
58	Finland	63	34
59	France	857	397
60	French Guiana	1	
61	Gabon	3	3
62	Gambia, The	1	
63	Georgia	21	12
64	Germany	740	407
65	Ghana	17	4
66	Gibraltar	1	
67	Greece	140	34
68	Greenland	1	
69	Guadeloupe	4	1
70	Guam	1	1
71	Guatemala	5	2
72	Guinea	5	3
73	Guinea-Bissau	3	3
74	Guyana	1	
75	Haiti	7	2
76	Honduras	4	1
77	Hong Kong	5	
78	Hungary	81	58
79	Iceland	15	8
80	India	136	69
81	Indonesia	8	7
82	Iran	61	20
83	Iraq	29	8
84	Ireland	73	29
85	Isle of Man	4	3
86	Israel	73	20
87	Italy	793	194
88	Jamaica	10	3
89	Japan	137	75
90	Jersey	1	
91	Jordan	7	3
92	Kazakhstan	10	6
93	Kenya	10	8
94	Korea, North	6	4
95	Korea, South	37	17
96	Kosovo	7	
97	Kuwait	3	2
98	Kyrgyzstan	5	4
99	Laos	1	1
100	Latvia	18	11
101	Lebanon	13	6
102	Lesotho	1	
103	Liberia	5	2
104	Libya	11	2
105	Lithuania	28	19
106	Luxembourg	8	4
107	Macedonia	15	3
108	Madagascar	2	1
109	Malawi	4	4
110	Malaysia	6	4
111	Maldives	3	1
112	Mali	8	4
113	Malta	3	2
114	Martinique	3	2
115	Mauritania	1	
116	Mauritius	1	1
117	Mexico	56	23
118	Micronesia, Federated States	1	1
119	Moldova	5	2
120	Monaco	4	1
121	Mongolia	8	1
122	Montenegro	10	4
123	Morocco	14	7
124	Mozambique	6	3
125	Namibia	2	2
126	Nauru	1	
127	Nepal	4	3
128	Netherlands	162	56
129	New Caledonia	2	
130	New Zealand	17	9
131	Nicaragua	5	5
132	Niger	1	1
133	Nigeria	23	6
134	Norway	59	33
135	Oman	2	1
136	Pakistan	28	13
137	Palestinian State	14	2
138	Panama	4	3
139	Paraguay	13	3
140	Peru	21	12
141	Philippines	19	16
142	Poland	167	114
143	Portugal	88	16
144	Puerto Rico	6	
145	Qatar	1	
146	Romania	50	26
147	Russia	369	240
148	Rwanda	1	1
149	Saint Kitts and Nevis	1	
150	Saint Lucia	2	2
151	Samoa	1	1
152	Sao Tome and Principe	1	1
153	Saudi Arabia	35	9
154	Senegal	10	2
155	Serbia	60	12
156	Seychelles	1	1
157	Sierra Leone	1	
158	Singapore	7	4
159	Slovakia	24	6
160	Slovenia	15	3
161	Solomon Islands	1	
162	Somalia	8	3
163	South Africa	43	22
164	South Sudan	1	1
165	Spain	298	77
166	Sri Lanka	6	5
167	Sudan	4	4
168	Suriname	5	2
169	Swaziland	1	
170	Sweden	135	61
171	Switzerland	102	56
172	Syria	19	2
173	Taiwan	10	4
174	Tajikistan	1	
175	Tanzania	3	3
176	Thailand	7	5
177	Timor-Leste	3	3
178	Togo	5	2
179	Tonga	2	1
180	Trinidad and Tobago	5	2
181	Tunisia	18	7
182	Turkey	184	35
183	Turkmenistan	3	1
184	Uganda	5	3
185	Ukraine	100	58
186	United Arab Emirates	5	4
187	United Kingdom	1,140	508
188	United States	2,291	1,221
189	Uruguay	23	7
190	Uzbekistan	9	1
191	Vanuatu	1	1
192	Venezuela	12	3
193	Vietnam	10	9
194	Virgin Islands	2	1
195	Yemen	6	2
196	Zambia	3	3
197	Zimbabwe	7	4
		10,773	4,886

Table S11 Number of people with articles in at least 26 Wikipedia language editions, by country.

	Language	Code	People (all years)	People (1800-1950)
1	**Afrikaans**	afr	6.94	4.14
2	**Albanian**	sqi	26.87	8.34
3	**Arabic**	ara	273.07	94.46
4	**Armenian**	hye	13.42	4.84
5	**Azerbaijani**	aze	25.79	9.74
6	**Basque**	eus	5.96	1.54
7	**Belarusian**	bel	5.15	2.34
8	**Bengali**	ben	18.86	12.45
9	**Bulgarian**	bul	22.35	6.18
10	**Catalan**	cat	51.06	13.09
11	**Chinese**	zho	115.6	44.24
12	**Czech**	ces	100.17	50.56
13	**Danish**	dan	100	39
14	**Dutch**	nld	226.86	81.26
15	**English**	eng	3300.8	1617.77
16	**Estonian**	est	10.1	6.06
17	**Filipino**	fil	19.22	16.22
18	**Finnish**	fin	57.46	31.01
19	**French**	fra	997.7	455.51
20	**Galician**	glg	20.86	5.39
21	**Georgian**	kat	14.91	8.52
22	**German**	deu	929.09	524.1
23	**Haitian**	hat	5.25	1.5
24	**Hebrew**	heb	58.4	16
25	**Hindi**	hin	55.95	28.39
26	**Hungarian**	hun	84.01	57.13
27	**Icelandic**	isl	15	8
28	**Italian**	ita	801.15	198.09
29	**Japanese**	jpn	137	75
30	**Kazakh**	kaz	6.3	3.78
31	**Kirghiz**	kir	3.23	2.59
32	**Korean**	kor	43	21
33	**Latvian**	lav	10.48	6.4
34	**Lithuanian**	lit	22.96	15.58
35	**Macedonian**	mkd	9.97	2
36	**Malay**	msa	15.99	12.56
37	**Malayalam**	mal	4.35	2.21
38	**Maltese**	mlt	2.71	1.8
39	**Maori**	mri	0.66	0.35
40	**Marathi**	mar	9.52	4.83
41	**Modern Greek**	ell	147.22	38.08
42	**Mongolian**	mon	7.2	0.9
43	**Norwegian**	nor	59	33
44	**Persian**	fas	42.83	15.6
45	**Polish**	pol	164.89	112.56
46	**Portuguese**	por	235.69	74.92
47	**Romanian**	ron	49.33	25.19
48	**Russian**	rus	429.38	272.91
49	**Serbo-Croatian**	hbs	152.84	36.92
50	**Sinhala**	sin	4.44	3.7
51	**Slovak**	slk	21.82	5.88
52	**Slovenian**	slv	13.66	2.73
53	**Spanish**	spa	774.64	305.48
54	**Swahili**	swa	12.4	10
55	**Swedish**	swe	138.47	62.87
56	**Tajik**	tgk	1.4	0.04
57	**Tamil**	tam	9.33	5.1
58	**Thai**	tha	7	5
59	**Turkish**	tur	164.86	33.64
60	**Turkmen**	tuk	3.21	1.22
61	**Ukrainian**	ukr	67.46	39.01
62	**Urdu**	urd	9.04	4.49
63	**Uzbek**	uzb	8.9	1.98
64	**Vietnamese**	vie	10.95	9.28

Table S12 Number of people with articles in at least 26 Wikipedia language editions, by language.

S5.3 Human Accomplishment

The book *Human Accomplishment: The Pursuit of Excellence in the Arts and Sciences, 800 B.C. to 1950*[27] 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.

To find the number of notable people for each language group, we converted countries of birth to languages as explained in **Section S5.2**. In most cases, we 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*. We replaced the first three with the specific places of birth for the respective people, as listed on *Wikipedia 26,* and converted them to languages based on their present-day countries. We 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 26 dataset, we increased the accuracy of the country-to-language mapping by selecting only the 1,655 people born between 1800 and 1950. Doing so also removed native speakers of Latin and Ancient Greek.

	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 S13 Number of people listed on human accomplishment, by country.

	Language	Code	People (all years)	People (1800-1950)
1	**Afrikaans**	afr	0.13	0.13
2	**Albanian**	sqi	0.88	0.47
3	**Arabic**	ara	86.05	14.05
4	**Basque**	eus	1.52	0.52
5	**Bengali**	ben	7.53	1.3
6	**Bulgarian**	bul	0.77	0.77
7	**Catalan**	cat	12.92	4.42
8	**Chinese**	zho	237.16	22.16
9	**Czech**	ces	45.79	26.71
10	**Danish**	dan	37	20
11	**Dutch**	nld	133.2	47.2
12	**English**	eng	788.1	466.26
13	**Finnish**	fin	5.47	4.56
14	**French**	fra	590.27	255.74
15	**Galician**	glg	5.32	1.82
16	**German**	deu	643.22	329.91
17	**Greek (Ancient)**	grc	134	
18	**Greek (Modern)**	ell	8.96	5.99
19	**Hindi**	hin	38.16	6.59
20	**Hungarian**	hun	20.5	17.62
21	**Icelandic**	isl	2	1
22	**Italian**	ita	393.22	60.14
23	**Japanese**	jpn	169	57
24	**Latin**	lat	55	
25	**Malayalam**	mal	2.98	0.51
26	**Maori**	mri	0.12	0.12
27	**Marathi**	mar	6.51	1.12
28	**Norwegian**	nor	23	22
29	**Polish**	pol	24.45	20.54
30	**Portuguese**	por	14.77	7.38
31	**Romanian**	ron	4.55	3.64
32	**Russian**	rus	134	118
33	**Serbo-Croatian**	hbs	11.61	8.11
34	**Slovak**	slk	4.12	3.8
35	**Slovenian**	slv	1.82	1.82
36	**Spanish**	spa	104.02	63.01
37	**Swahili**	swa	0.8	0.8
38	**Swedish**	swe	44.33	21.27
39	**Tamil**	tam	5.49	0.94
40	**Turkish**	tur	1.81	1.19
41	**Ukrainian**	ukr	0.04	0.04
42	**Urdu**	urd	4.65	0.8
43	**Vietnamese**	vie	0.04	0.04

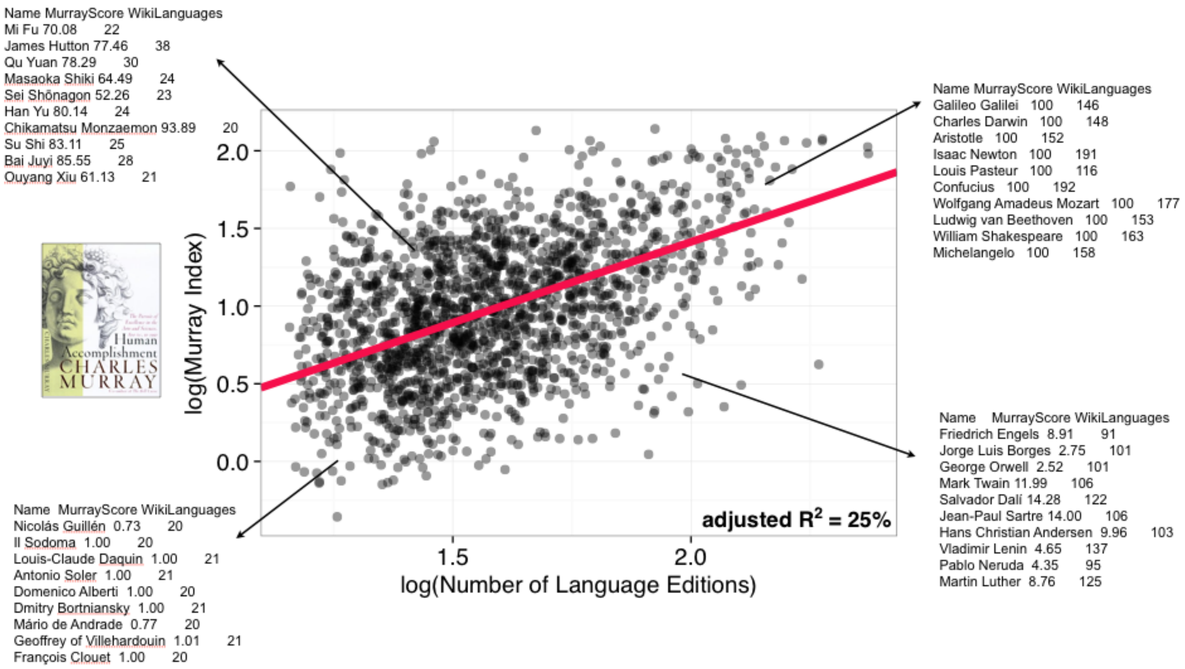
Table S14 Number of people listed on human accomplishment, by language.

S5.4 Comparison of the famous people datasets

The two datasets we use—*Wikipedia 26* and *Human Accomplishment*—were compiled in different ways. Wikipedia is written by a large number of volunteers with different backgrounds from all over the world, while Human Accomplishment is the work of a single author, the American political scientist Charles Murray. Naturally, both sources exhibit certain biases despite the efforts taken by their authors.

To understand these biases, we compared the cultural significance attributed by each dataset to the listed individuals. We define the cultural significance of a person as the number of languages in which his/her Wikipedia biography is available (for entries on Wikipedia 26), or the score that Murray gave this individual (*Human Accomplishment* entries are given a score from 1 to 100 based on their contribution in their respective field or fields). **Figure S5** shows the correlation between these two measurements. One notable observation is that the cultural contribution the Charles Murray attributes to people born in Asia (measured by their score on his list) is higher than their cultural contribution according to *Wikipedia 26* (measured by the number of languages in which a Wikipedia biography is available). Murray is also less likely than Wikipedia to acknowledge the contribution of left-wing liberals.

The moderate correlation (R^2=0.25) shows that using these two lists of famous individuals provides a more balanced perspective than the exclusive use of Wikipedia. While the two datasets are substantially different, there is a consistent correlation between the number of famous people in a language according to either dataset and the centrality of that language, attesting to the robustness of our method.



Name	MurrayScore	WikiLanguages
Mi Fu	70.08	22
James Hutton	77.46	38
Qu Yuan	78.29	30
Masaoka Shiki	64.49	24
Sei Shōnagon	52.26	23
Han Yu	80.14	24
Chikamatsu Monzaemon	93.89	20
Su Shi	83.11	25
Bai Juyi	85.55	28
Ouyang Xiu	61.13	21

Name	MurrayScore	WikiLanguages
Galileo Galilei	100	146
Charles Darwin	100	148
Aristotle	100	152
Isaac Newton	100	191
Louis Pasteur	100	116
Confucius	100	192
Wolfgang Amadeus Mozart	100	177
Ludwig van Beethoven	100	153
William Shakespeare	100	163
Michelangelo	100	158

Name	MurrayScore	WikiLanguages
Nicolás Guillén	0.73	20
Il Sodoma	1.00	20
Louis-Claude Daquin	1.00	21
Antonio Soler	1.00	21
Domenico Alberti	1.00	20
Dmitry Bortniansky	1.00	21
Mário de Andrade	0.77	20
Geoffrey of Villehardouin	1.01	21
François Clouet	1.00	20

Name	MurrayScore	WikiLanguages
Friedrich Engels	8.91	91
Jorge Luis Borges	2.75	101
George Orwell	2.52	101
Mark Twain	11.99	106
Salvador Dalí	14.28	122
Jean-Paul Sartre	14.00	106
Hans Christian Andersen	9.96	103
Vladimir Lenin	4.65	137
Pablo Neruda	4.35	95
Martin Luther	8.76	125

Figure S5 Correlation of the *Wikipedia 26* and *Human Accomplishment* datasets

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Number of illustrious people born 1800-1950 per language, based on inclusion in *Human Accomplishment.* English excluded.						
$\log_{10}$(Population)	0.801***				0.972***	0.292	0.302*
	(0.124)				(0.189)	(0.222)	(0.113)
$\log_{10}$(GDP per capita)	1.898***				2.348***	0.616	0.602*
	(0.287)				(0.472)	(0.549)	(0.279)
EV centrality [Twitter]		0.414**			-0.162		
		(0.112)			(0.136)		
EV centrality [Wikipedia]			0.985***			0.696*	
			(0.117)			(0.263)	
EV centrality [book trans.]				0.961***			0.734***
				(0.083)			(0.119)
(Intercept)	-8.304***	2.030***	2.460***	2.824***	-10.862***	-1.045	-0.647
	(1.363)	(0.271)	(0.174)	(0.157)	(2.532)	(3.004)	(1.518)
Observations	28	28	28	28	28	28	28
p-value	0	0.001	0	0	0	0	0
R-squared	0.679	0.344	0.733	0.838	0.697	0.751	0.875
Adjusted R-squared	0.653	0.319	0.723	0.832	0.659	0.72	0.86

***,**,* significant at 0.1%, 1% and 5% levels, respectively. Standard errors in parentheses.
Only languages with at least one illustrious person are included.

Table S16 Regression table for *Human Accomplishment* considering people born between 1800-1950 (see Table 2A in the main text) but excluding English