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

Shahar Ronen[a], Bruno Gonçalves[b,c,d], Kevin Z. Hu[a], Alessandro Vespignani[b], Steven Pinker[e], and César A. Hidalgo[a,1]

[a]Macro Connections, Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139; [b]Department of Physics, Northeastern University, Boston, MA 02115; [c]Aix-Marseille Université, CNRS, CPT, UMR 7332, 13288 Marseille, France; [d]Université de Toulon, CNRS, CPT, UMR 7332, 83957 La Garde, France; and [e]Department of Psychology, Harvard University, Cambridge, MA 02138

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved November 13, 2014 (received for review June 11, 2014)

Languages vary enormously in global importance because of historical, demographic, political, and technological forces. However, beyond simple measures of population and economic power, there has been no rigorous quantitative way to define the global influence of languages. Here we use the structure of the networks connecting multilingual speakers and translated texts, as expressed in book translations, multiple language editions of Wikipedia, and Twitter, to provide a concept of language importance that goes beyond simple economic or demographic measures. We find that the structure of these three global language networks (GLNs) is centered on English as a global hub and around 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 it exhibits a strong correlation with two independent measures of the number of famous people born in the countries associated with that language. These results suggest that the position of a language in the GLN contributes to the visibility of its speakers and the global popularity of the cultural content they produce.

networks | languages | culture | digital humanities | fame

Of the thousands of languages that have ever been spoken, only a handful have become influential enough to be considered global languages. However, how do we measure the global influence of a language? 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. Several studies have relied on measures that proxy the global influence of a language using the population and wealth of its speakers (1–4). While wealth and population approximate a language's influence, as the dissemination of a language has historically required a strong power base (5), such measures fail to capture the global influence of a language: often the speakers of a language, and their wealth, are locally concentrated, making the language locally influential rather than globally influential.

An alternative method to measure the global influence of a language is to focus on who speaks that language, and in particular, on how connected the speakers of that language are. In the words of linguist David Crystal, "Why a language becomes a global language has little to do with the number of people who speak it. It is much more to do with who those speakers are." (5) In the past, Latin was the pan-European language, not because it was the mother tongue of most Europeans, but because it was the language of the Roman Empire and later the language of the Catholic Church, scholars, and educators (5). The use of Latin by well-connected elites set it apart from other languages and helped Latin endure as a universal language for more than 1,000 years.

However, can we use these ideas to identify which modern languages are globally influential? If global languages are those connecting international elites, then we can identify the global languages associated with particular elites by mapping their networks of multilingual coexpressions. Examples of multilingual coexpressions include book translations, edits to multiple language editions of Wikipedia, and posting short messages on Twitter ("tweets") in multiple languages. These coexpressions define networks (Fig. 1) that—even though not representative of the world's general population—represent a coarse map of the links connecting the elites that participate of these three important global forums, as social connections often require a shared language.

In this paper, we map the global language networks (GLNs) expressed in three large records of linguistic expression, and use the structure of these networks to determine the degree to which each language is global. First, we look at a collection of more than 2.2 million book translations compiled by UNESCO's *Index Translationum* project. This dataset allows us to map the network of book translations, which are produced by individuals with a high literary capacity (authors and professional translators) and are shaped by market forces, such as the demand for books in different languages. Each translation from one language to another forms a connection. Next, we map the network of linguistic coexpressions expressed by the community of digitally engaged knowledge specialists that edit Wikipedia. Here, two languages are connected when users that edit an article in one Wikipedia language edition are significantly more likely to also edit an article in another language edition. Finally, we map the network of linguistic coexpressions expressed in Twitter. Here, two languages are connected when users that tweet in a language are also significantly more likely to tweet in another language.

These three networks allow us to map the paths of direct and indirect communication between speakers from different languages. Our method formalizes the intuition that certain

Significance

People have long debated about the global influence of languages. The speculations that fuel this debate, 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 a language's global influence based on its position 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.

Author contributions: S.R. and C.A.H. designed and performed research; S.R., B.G., and K.Z.H. retrieved data; S.R., K.Z.H., and C.A.H. analyzed data; B.G., A.V., and S.P. advised on analytic methods; and S.R., S.P., and C.A.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: Visit language.media.mit.edu for datasets and additional visualizations.

[1]To whom correspondence should be addressed. Email: hidalgo@mit.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1410931111/-/DCSupplemental.

Downloaded from https://www.pnas.org by 37.19.197.239 on October 16, 2023 from IP address 37.19.197.239.

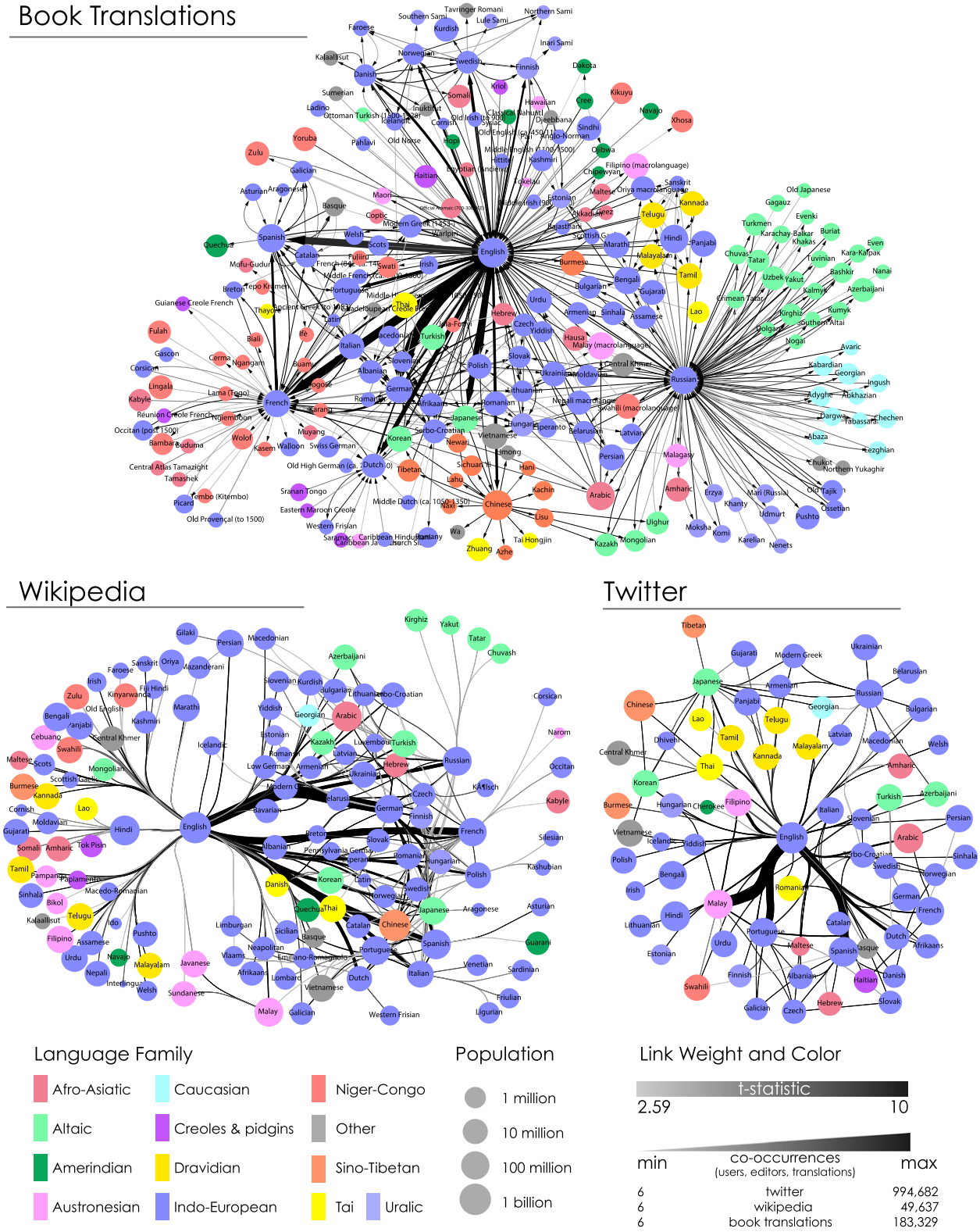


Fig. 1. Visualizations of the three GLNs. The three GLNs contain all language connections that involve at least six users (Twitter and Wikipedia) or six translations and that are significant with $P < 0.01$.

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 an Englishman through bilingual speakers of English and Spanish. An idea conceived by a Vietnamese speaker, 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

Downloaded from https://www.pnas.org by 37.19.197.239 on October 16, 2023 from IP address 37.19.197.239.

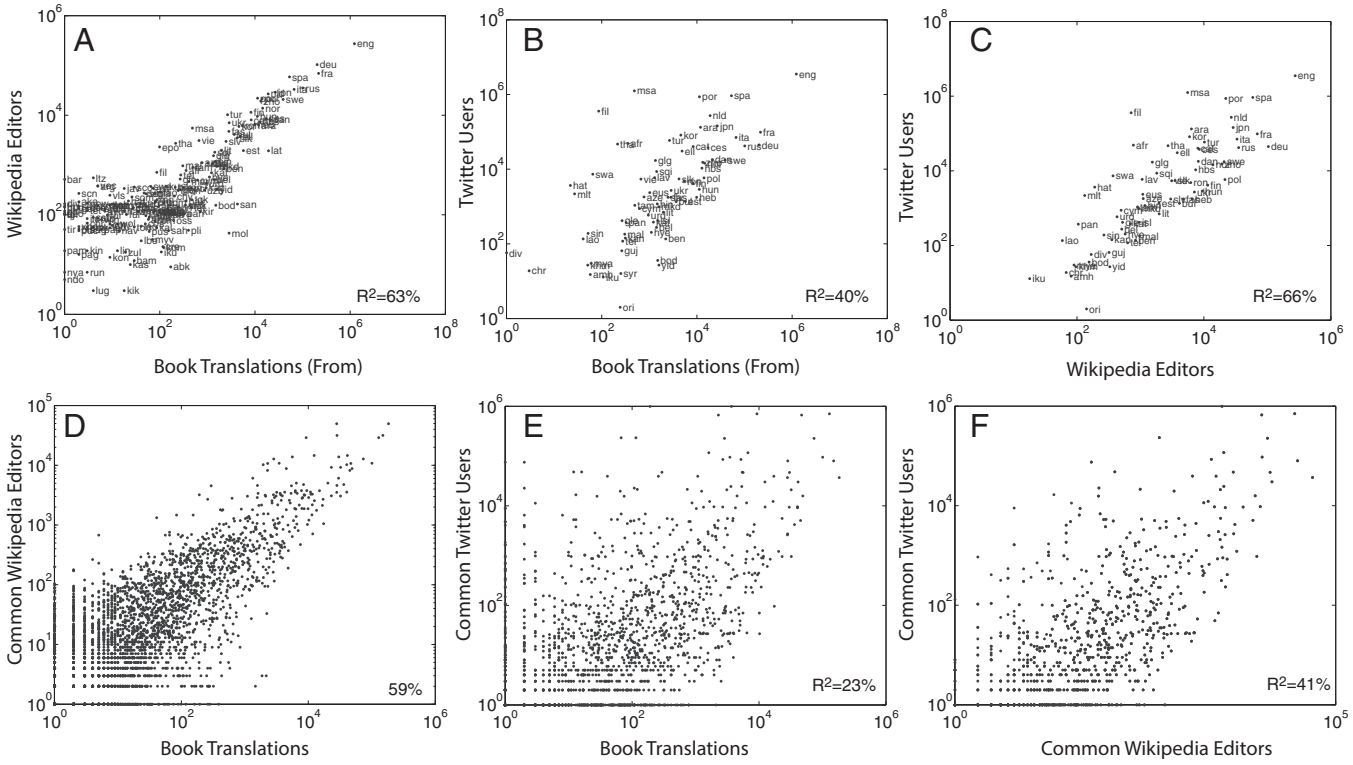


Fig. 2. Similarity of the three independent datasets we use for mapping the GLNs. The top row shows the correlation between the number of expressions for each language across the three datasets: (*A*) Wikipedia editors in a language and book translations from a language; (*B*) Twitter users in a language and book translations from a language; and (*C*) Twitter users and Wikipedia editors. The bottom row shows the correlation between the number of coexpressions 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* and *E*, we symmetrized the book translation network by considering the average of translations from and to a language.

Mapudungun. In both cases, however, English and Spanish are still involved in the flow of information, indicating that they act as global languages. In the first example, Spanish and English have a direct involvement because communication is flowing among their speakers. In the latter case, the involvement is indirect and emerges from the lack of speakers that can communicate in both Vietnamese and Mapudungun. These indirect connections make multilingual speakers of global languages globally influential, as they mediate the flow of information not only among each other, but also, among people with whom they do not share a language (6).

Influence of Global Languages

The position of a language in the global language network is expected to affect the visibility of the content produced by the speakers of a language and also the flow of information among the speakers of different languages. Intuitively, better connected languages should increase the visibility of the content produced by their speakers: if information radiates from the more connected to the less connected languages, it will be easier for an English speaker than for a Nepali speaker to become globally famous.

The position of a language in the GLN, however, also affects the flow of information that is not produced in that language. The degree to which a language is global generates incentives to create content in that language and to translate content produced in less-connected languages into that language. For example, a reporter wishing to disseminate news about a major event around the world has an incentive to translate the news into a global language or to report the news in a global language to begin with (7). The position of a language in the global language network therefore affects the diversity of international information available to its speakers, the speed with which they receive international information, and their ability to disseminate information to the speakers of other languages.

Creating a map linking languages that are likely to be cospoken is an important step for understanding the relevance of the global language network in the diffusion of information. In this paper, we map three GLNs to develop a metric of the degree to which a language is global and validate these networks by showing that the centrality of a language in the GLN is strongly correlated with the number of famous people born in that language.

Data

There is no single 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 (8).

Going forward, we note that the resulting networks represent patterns of linguistic coexpression 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

Downloaded from https://www.pnas.org by 37.19.197.239 on October 16, 2023 from IP address 37.19.197.239.

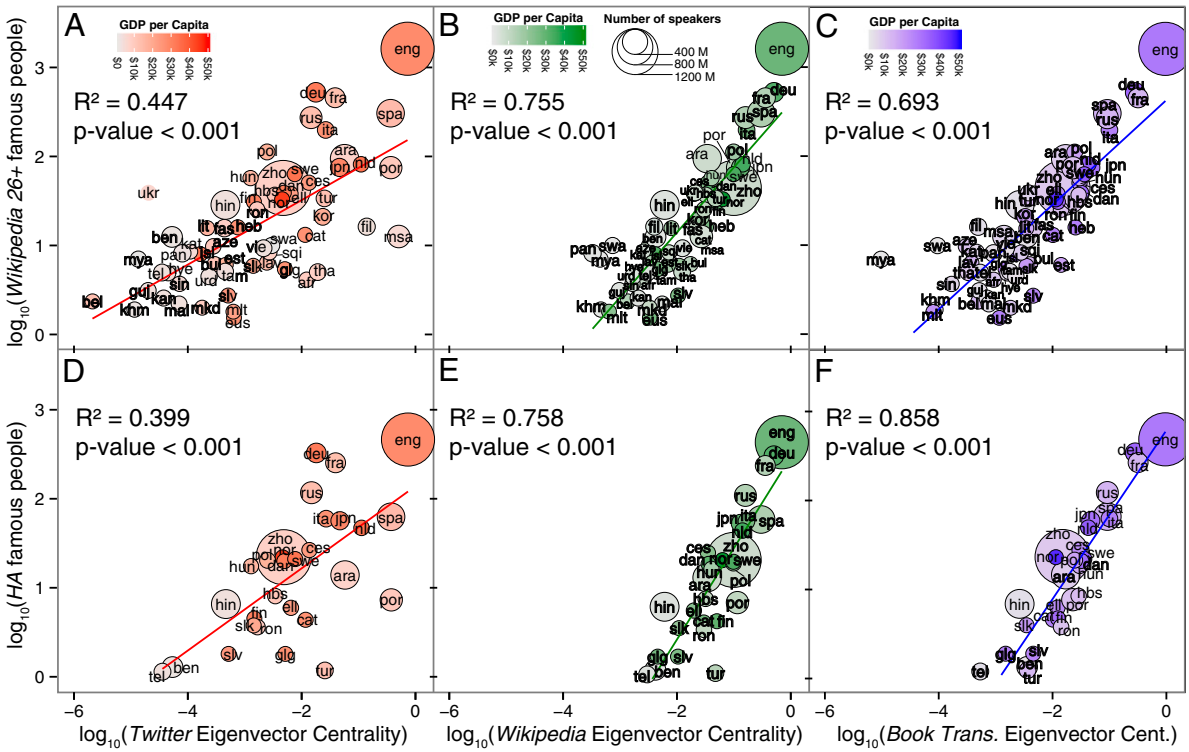


Fig. 3. The position of a language in the GLN and the global impact of its speakers. Top row shows the number of people per language (born 1800–1950) with articles in at least 26 Wikipedia language editions as a function of their language's eigenvector centrality in the (*A*) Twitter GLN, (*B*) Wikipedia 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. All subplots report the adjusted R^2.

Baidu Baike, do not represent resources that can be used to map connections between global languages.

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

We compiled our Twitter dataset from more than 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 (9), 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% (*SI Appendix*). Our final dataset consists of nearly 550 million tweets in 73 languages generated by more than 17 million unique users, which represented more than 10% of Twitter's active users at the time the data were collected. Two languages are connected when users that tweet in one language are significantly more likely to also tweet in the other language (Eq. **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 *SI Appendix*, the dataset contains 382 million edits in 238 languages by 2.5 million unique editors. Here, two languages are connected when users that edit an article in one Wikipedia language edition are significantly more likely to also edit an article in the edition of the other language (Eq. **1**).

The IT dataset consists of 2.2 million translations of printed books published between 1979 and 2011 in 150 countries and more than a thousand languages (*SI Appendix*). The dataset records translations rather than books, so it does not list books that have not been translated. Moreover, the IT dataset 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, at the time we retrieved the data, the IT recorded 15 translations of *The Adventures of Tom Sawyer* to Catalan, of which only 13 were translated directly from English; the two other translations came from Spanish and Galician versions. This characteristic of the dataset allows us to identify languages that serve as intermediaries for translations.

In all three cases, we collapsed mutually intelligible languages following the ISO 639-3 standard (10). 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 *SI Appendix*.

Results

Mapping a Global Language Network. We begin our construction of these three GLNs by identifying the links that are statistically significant with respect to the population of speakers expressed in each dataset. By definition, a statistically 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 $[P(i,j) > P(i)P(j)]$. To determine the strength and significance of each connection, we use the standard methods of ϕ correlation and t statistic, which are defined as follows.

Downloaded from https://www.pnas.org by 37.19.197.239 on October 16, 2023 from IP address 37.19.197.239.

A

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Number of famous people born 1800-1950 per language, based on having biographies in at least 26 Wikipedia 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 famous person are included.

B

Famous 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

Famous 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

Fig. 4. 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 about which there are articles in at least 26 Wikipedia language editions as a function of the language's GDP per capita, population, and eigenvector (EV) centrality in each of GLNs. Cultural production rankings: the (*B*) countries and (*C*) languages with 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 *SI Appendix* for the full lists.

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

$$\phi_{ij} = \frac{M_{ij}N - M_iM_j}{\sqrt{M_iM_j(N-M_i)(N-M_j)}}, \quad [1]$$

where M_i represents the number of multilingual users (or translations) expressed in language i ($M_i = \sum_j M_{ij}$), and N represents the total number of users or translations in the dataset. ϕ_{ij} is positive for pairs of languages that co-occur more often than expected based on their representation in the dataset alone and is negative otherwise. To assess the statistical significance of these correlations, we use the t statistic, which is given by

$$t_{ij} = \frac{\phi_{ij}\sqrt{D-2}}{\sqrt{1-\phi_{ij}^2}}, \quad [2]$$

where $D-2$ represents the degrees of freedom of the correlation. Here we consider $D = \max(M_i, M_j)$ as it provides more stringent criteria for finding a correlation than using $D = N$.

Finally, we construct our network by considering only links that are statistically significant with $P < 0.01$ ($t_{ij} > 2.59$ for $D > 20$; one-tailed). Also, we consider only links for which $M_{ij} \geq 6$ to avoid false positives that could emerge due to small statistics. In sum, we discard links when either $t_{ij} \leq 2.59$ (they are statistically insignificant) or $M_{ij} < 6$ (the sample size is too small). We note that, by definition, a null model network would contain no links because none of the links of a null model network would satisfy the statistical significance condition ($t_{ij} > 2.59$). We also note that the book translation GLN is directed, unlike the Twitter and Wikipedia GLN, because we know the source and target of each translation. For the book translation network we assessed the significance of each directed link separately and kept zero, one, or both directed links between any given pair of languages.

The Structure of Three GLNs. To understand the relative importance of each language, we begin by visualizing the three GLNs (Fig. 1). In this visualization, each node represents a language. Node sizes are proportional to the number of speakers (native plus nonnative) of each language (11). 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 number of co-occurrences or translations (M_{ij}).

The three GLNs share a number of features, even though they capture information about the linguistic expression of different communities. First, the representation or number of expression of each language—number of Twitter users, Wikipedia editors, or translations from a language—correlates strongly across the three datasets (Fig. 2 *A*–*C*). Moreover, the coexpressions—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 (Fig. 2 *D*–*F*). This correlation means that a language with a high or low coexpression to another language in one GLN is likely to have a high or low coexpression with that same language in the other GLNs. The positive correlations of expressions and of coexpressions across GLNs suggest that the three networks, although representative of their own respective populations, are similar in terms of the size of the populations observed in them and the strength of their links.

It is interesting to note that the similarities observed between the networks appear to conform to their gradient of formality, defined in terms of the literary capacity required by a speaker to participate of each of these global forums. The book translation network is the most formal (as it involves published authors and professional translators) and Twitter is the least formal (as it consists of short, instant messages that anyone with internet access can write). Wikipedia takes the middle ground between Twitter and book translations in terms of formality, and its GLN takes the middle ground also in terms of similarity. The R^2 between the representation of a language in the Wikipedia GLN and in the book translation GLN is 63% and between the Wikipedia GLN and the Twitter GLN is 66%. For coexpressions, the R^2 is 41% and 63%, respectively. By contrast, the similarity between the representation of languages in Twitter and book translations is $R^2 = 40\%$, and their similarity in coexpression is only $R^2 = 23\%$, meaning that the Twitter and book translation

Downloaded from https://www.pnas.org by 37.19.197.239 on October 16, 2023 from IP address 37.19.197.239.

A

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Number of famous 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 famous person are included.

B

Famous 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

Famous 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

Fig. 5. 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 that are listed in *Human Accomplishment* (*HA*) as a function of the language's GDP per capita, population, and eigenvector (EV) centrality in each of GLNs. Cultural production rankings: the (*B*) countries and (*C*) languages with the largest number of people on the HA list. For the full lists, see *SI Appendix*.

networks are the most dissimilar. Finally, we note that compared with the book translation dataset, the two digital datasets (Twitter and Wikipedia) exhibit a larger share of languages associated with developing countries—such as Malay, Filipino, and Swahili—which could indicate that these less formal channels of communication are more inclusive of the populations of developing countries than written books.

The Position of Languages in the GLN. 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 (12); for other centrality measures, see *SI Appendix*. Eigenvector centrality (which is also the basis for Google's PageRank algorithm) 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 (data available at pantheon.media.mit.edu). The 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, and 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 *SI Appendix* 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 (13). We distributed the contribution of the 1,655 people on this list born between 1800 and 1950 across different languages using the same method used for the Wikipedia dataset (*SI Appendix*).

SI Appendix, Fig. S5 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 ($R^2 = 0.25$, $P \ll 0.001$). 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 colinearity of the data.

Fig. 3 *A–C* shows 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 are associated with one or more famous character. The regression table in Fig. 4 compares the effects of several independent variables: the combined effect of its population and income (combined) and the effect eigenvector centrality in each of the GLNs. The variables are introduced sequentially. We use gross domestic product (GDP) to indicate income; see *SI Appendix* for an explanation of the method used to calculate the population and GDP of a language.

With the exception of the Twitter dataset, the correlation between the number of famous people and the eigenvector centrality of a language is similar to or higher than the correlation observed between the number of famous people and the income and population of the language combined using both nested models (column 1 vs. 6 and 7 and columns 3 and 4 vs.

Downloaded from https://www.pnas.org by 37.19.197.239 on October 16, 2023 from IP address 37.19.197.239.

6 and 7) and nonnested statistical models (column 1 vs. columns 3 and 4). For nonnested models, we used a Clarke test and found that eigenvector centrality is a significantly better correlate of fame than the combination of population and income, with $P < 0.01$ for almost all combinations. The one exception is the combination of Wikipedia fame and the eigenvector centrality in the book translation GLN, for which both regressions are statistically equivalent. For nested models, we estimate the semipartial correlation and the F test. The semipartial correlation is 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. For the Wikipedia fame dataset, we find that the percentage of the variance in the number of famous people explained by the centrality of a language in the Wikipedia and book translation GLNs are, respectively, 7.5% ($F = 22.97$, $P < 0.001$) and 7.7% ($F = 23.48$, $P < 0.001$) after the effects of income and population have been taken into account. In contrast, the semipartial contributions of income and population together are 5.1% ($F = 7.74$, $P < 0.001$) when measured against the Wikipedia GLN, and 11.3% ($F = 17.32$, $P < 0.001$) when measured against the book translation GLN.

Figs. 3 *D–F* and 5 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 (Fig. 5), 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% ($F = 7.59$, $P = 0.01$) and 16.1% ($F = 37.98$, $P < 0.001$) at the margin, as measured by the semipartial correlation. The semipartial contribution of income and population in this case is much lower, being only 2.3% ($F = 1.43$, $P = 0.26$) and 2.7% ($F = 3.13$, $P = 0.06$) when measured, respectively, against the Wikipedia GLN and book translation GLN.

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, because 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 previously unidentified characterization of a language's global importance. The GLNs, mapped from millions of online and printed linguistic expressions, reveal that the world's languages exhibit a hierarchical structure dominated by a central hub, English, and a halo of intermediate hubs, which include other global languages such as German, French, and Spanish. Although 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, which had been identified as an obstacle to the dissemination of outside knowledge into the Arab world (14), is indicated by our book translation GLN and matched by the peripheral position of Arabic in the Twitter and Wikipedia GLNs.

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, although 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, because a global language also links distant languages and not just local or regional ones.

The structure of the three GLNs documented here also raises important questions involving the dynamics of the networks observed. Future assessments of temporal changes in the structure of the GLNs (which will be possible as data for a longer period of time becomes available) 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 (15).

1. Davis M (2003) *GDP by Language*. Available at www.unicode.org/notes/tn13. Accessed June 30, 2012.
2. Ostler N (2005) *Empires of the Word: a Language History of the World* (HarperCollins, New York).
3. Pimienta D, Prado D, Blanco Á (2009) *Twelve Years of Measuring Linguistic Diversity in the Internet: Balance and Perspectives*. Available at www.ifap.ru/pr/2010/n100305c.pdf. Accessed December 14, 2012.
4. Weber G (1997) The world's 10 most influential languages. *Language Today* 2(3): 12–18.
5. Crystal D (2003) *English as a Global Language* (Cambridge Univ Press, Cambridge, UK).
6. Chambers JK (2009) *Sociolinguistic Theory: Linguistic Variation and Its Social Significance* (Wiley-Blackwell, Oxford).
7. Zuckerman E (2013) *Rewire: Digital Cosmopolitans in the Age of Connection* (WW Norton, New York).
8. UNESCO, *Index Translationum: World Bibliography of Translation*. Available at www.unesco.org/xtrans/bsform.aspx. Accessed July 22, 2012.
9. McCandless M (2011) *Chromium Compact Language Detector*. Available at code.google.com/p/chromium-compact-language-detector/. Accessed November 6, 2011.
10. International SIL (2007) *ISO 639-3 Registration Authority*. Available at www.sil.org/iso639-3. Accessed June 14, 2012.
11. Zachte E (2012) *Wikimedia Statistics*. Available at: stats.wikimedia.org/EN/Sitemap.htm. Accessed June 14, 2012.
12. Bonacich P (1987) Power and centrality: A family of measures. *Am J Sociol* 92(5): 1170–1182.
13. Murray CA (2003) *Human Accomplishment: The Pursuit of Excellence in the Arts and Sciences, 800 B.C. to 1950* (HarperCollins, New York).
14. United Nations Development Programme (2003) *Arab Human Development Report 2003: Building a Knowledge Society*. Available at www.arab-hdr.org/publications/other/ahdr/ahdr2003e.pdf. Accessed January 31, 2013.
15. Abrams DM, Strogatz SH (2003) Linguistics: Modelling the dynamics of language death. *Nature* 424(6951):900.

Downloaded from https://www.pnas.org by 37.19.197.239 on October 16, 2023 from IP address 37.19.197.239.