Supplementary Information Appendix for *Links that speak: The global language network and its association with global fame*

Shahar Ronen, Bruno Gonçalves, Kevin Z. Hu, Alessandro Vespignani, Steven Pinker, César A. Hidalgo

*Supplementary online material (SOM) and additional visualizations are available on [http://language.media.mit.edu](http://language.media.mit.edu)*

Table of Contents

S1 Data ......................................................................................................................................................... 2
  S1.1 Twitter ............................................................................................................................................... 2
  S1.2 Wikipedia .......................................................................................................................................... 4
  S1.3 Book translations .......................................................................................................................... 7
S2 Language notation and demographics ................................................................................................. 8
  S2.1 Notation ........................................................................................................................................... 8
  S2.2 Population ...................................................................................................................................... 9
  S2.3 Language GDP ............................................................................................................................ 10
S3 Additional calculations .......................................................................................................................... 11
S4 Language centrality: Eigenvector centrality vs. betweenness centrality ............................................ 12
S5 Famous people per language .............................................................................................................. 14
  S5.1 Associating a famous person with languages .............................................................................. 14
  S5.2 Wikipedia ..................................................................................................................................... 18
  S5.3 Human Accomplishment ........................................................................................................... 22
  S5.4 Comparison of the famous people datasets .............................................................................. 24
References for the SI Appendix .................................................................................................................. 26
**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 (1).

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 (2). For example, Twitter users in the United States are younger and hold more liberal opinions than the general public (3).

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) (4), which was chosen for its wide language support and its relatively accurate detection of short messages (5, 6). However, any automated language detection is prone to errors (7), 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).

![Graph A](image1.png) ![Graph B](image2.png)

**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 (8) produced by CLD to three-letter ISO 639-3 codes (9), 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 (10). 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 site. 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.

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

Table S1 Statistics for the twenty languages with the most tweets in our Twitter dataset. The full table is available on 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 (11). 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 edits). However, this user contributed to ten additional language editions, in particular Somali (149 edits) and Japanese (58 edits). 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 site. We used this dataset to generate the Wikipedia GLN shown in Figure 1 of the main section.

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

Table S2 Statistics for the twenty languages with the most edits in our Wikipedia dataset. The full table is available on the SOM site.
S1.3 Book translations

The Index Translationum is an international bibliography of book translations maintained by UNESCO (12). 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 (13).

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 (9)). 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.

<table>
<thead>
<tr>
<th>#</th>
<th>Language</th>
<th>Code</th>
<th>Translations from</th>
<th>Translations to</th>
<th>Total translations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>English</td>
<td>eng</td>
<td>1,225,237</td>
<td>146,294</td>
<td>1,371,531</td>
</tr>
<tr>
<td>2</td>
<td>German</td>
<td>deu</td>
<td>201,718</td>
<td>292,124</td>
<td>493,842</td>
</tr>
<tr>
<td>3</td>
<td>French</td>
<td>fra</td>
<td>216,624</td>
<td>238,463</td>
<td>455,087</td>
</tr>
<tr>
<td>4</td>
<td>Spanish</td>
<td>spa</td>
<td>52,955</td>
<td>228,910</td>
<td>281,867</td>
</tr>
<tr>
<td>5</td>
<td>Russian</td>
<td>rus</td>
<td>101,395</td>
<td>82,772</td>
<td>184,167</td>
</tr>
<tr>
<td>6</td>
<td>Japanese</td>
<td>jpn</td>
<td>26,921</td>
<td>130,893</td>
<td>157,814</td>
</tr>
<tr>
<td>7</td>
<td>Dutch</td>
<td>nld</td>
<td>18,978</td>
<td>111,371</td>
<td>130,349</td>
</tr>
<tr>
<td>8</td>
<td>Italian</td>
<td>ita</td>
<td>66,453</td>
<td>59,830</td>
<td>126,283</td>
</tr>
<tr>
<td>9</td>
<td>Swedish</td>
<td>swe</td>
<td>39,192</td>
<td>71,688</td>
<td>110,880</td>
</tr>
<tr>
<td>10</td>
<td>Polish</td>
<td>pol</td>
<td>14,104</td>
<td>76,720</td>
<td>90,824</td>
</tr>
<tr>
<td>11</td>
<td>Portuguese</td>
<td>por</td>
<td>11,390</td>
<td>74,721</td>
<td>86,111</td>
</tr>
<tr>
<td>12</td>
<td>Danish</td>
<td>dan</td>
<td>21,239</td>
<td>64,799</td>
<td>86,038</td>
</tr>
<tr>
<td>13</td>
<td>Czech</td>
<td>ces</td>
<td>17,202</td>
<td>64,442</td>
<td>81,644</td>
</tr>
<tr>
<td>14</td>
<td>Chinese</td>
<td>zho</td>
<td>13,337</td>
<td>62,650</td>
<td>75,987</td>
</tr>
<tr>
<td>15</td>
<td>Hungarian</td>
<td>hun</td>
<td>11,256</td>
<td>54,989</td>
<td>66,245</td>
</tr>
<tr>
<td>16</td>
<td>Norwegian</td>
<td>nor</td>
<td>14,530</td>
<td>45,923</td>
<td>60,453</td>
</tr>
<tr>
<td>17</td>
<td>Serbo-Croatian</td>
<td>hbs</td>
<td>12,743</td>
<td>45,036</td>
<td>57,779</td>
</tr>
<tr>
<td>18</td>
<td>Finnish</td>
<td>fin</td>
<td>8,296</td>
<td>46,271</td>
<td>54,567</td>
</tr>
<tr>
<td>19</td>
<td>Modern Greek</td>
<td>ell</td>
<td>4,862</td>
<td>27,422</td>
<td>32,284</td>
</tr>
<tr>
<td>20</td>
<td>Bulgarian</td>
<td>bul</td>
<td>3,667</td>
<td>25,742</td>
<td>29,409</td>
</tr>
</tbody>
</table>

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

Languages belong to language families (14). We mapped languages to language families using the hierarchy in Ethnologue (15) 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 (16).
**S2.2 Population**

We use language speaker estimates from the June 14, 2012 version of Wikipedia Statistics page (17). 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 languages in our datasets.

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

Table S4 Macrolanguages for which the estimated number of speakers is not the sum of the estimates for the individual languages that constitute them.
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 (18). 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 (18) with a few additions from the CIA World Factbook (19)). Also, given a language $l$, let $N_{lc}$ be the number of native speakers of $l$ in country $c$, ...
obtained from Ethnologue (15) and The World Factbook (19). We calculated $N_{lc}$ using the language demographics listed in Table S6. Thus, $G_i$, the GDP per capita for $i$ is

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

The GDP per capita values in Table S5 are approximate, 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 site.

**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: Eigenvector centrality vs. betweenness centrality

In this section we compare two measures of centrality, eigenvector centrality (the metric used in the main text) and 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. 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. 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 shows 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. A table with eigenvector and betweenness centralities of each language in the Twitter, Wikipedia and book translation GLNs is available on the SOM site.

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 to 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). 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.
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 (21), 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 famous 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 (15, 19). 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 Table S6 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.
<table>
<thead>
<tr>
<th></th>
<th>Country</th>
<th>Language demographics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Afghanistan</td>
<td>Persion 50%, Punsho 35%, Uzbek 6%, Turkmen 5%</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Albania</td>
<td>Albanian 95%, Greek (Modern) 3%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Algeria</td>
<td>Arabic 80%, French 20%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Andorra</td>
<td>Catalan 40%, Spanish 35%, Portuguese 15%, French 5%</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Angola</td>
<td>Portuguese 70%</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Argentina</td>
<td>Spanish 98%</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Armenia</td>
<td>Armenian 97.7%, Russian 0.9%</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Aruba</td>
<td>Papiamento 66.3%, Spanish 12.6%, English 7.7%, Dutch 5.8%</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Australia</td>
<td>English 78.5%, Chinese 2.5%, Italian 1.6%, Greek (Modern) 1.3%, Arabic 1.2%, Vietnamese 1%</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Austria</td>
<td>German 88.6%, Serbo-Croatian 3.8%, Turkish 2.3%</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Azerbaijan</td>
<td>Azerbaijani 90.3%, Lezginian 2.2%, Russian 1.8%, Armenian 1.5%</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Bahamas, The</td>
<td>English 100%</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Bahrain</td>
<td>Arabic 100%</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Bangladesh</td>
<td>Bengali 98%</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Barbados</td>
<td>English 100%</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Belarus</td>
<td>Russian 70.2%, Belarusian 23.4%</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Belgium</td>
<td>Dutch 60%, French 40%</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Belize</td>
<td>English 41%, Spanish 32%</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Benin</td>
<td>French 49%, Fon 39%, Yoruba 12%</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Bermuda</td>
<td>English 100%</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Bhutan</td>
<td>Tshangka 28%, Dzonghka 24%, Nepali 22%</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Bolivia</td>
<td>Spanish 60.7%, Quechua 21.2%, Aymara 14.6%</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Bosnia and Herzegovina</td>
<td>Serbo-Croatian 100%</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Botswana</td>
<td>Tswana 78.2%, Kalanga 7.9%, English 2.1%</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Brazil</td>
<td>Portuguese 100%</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Brunei</td>
<td>Malay 100%</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Bulgaria</td>
<td>Bulgarian 76.5%, Turkish 8.2%, Romany 3.8%</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Burkinsa Faso</td>
<td>French 100%</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>Burma</td>
<td>Burmese 100%</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>Burundi</td>
<td>French 50%, Rundi 50%</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>Cambodia</td>
<td>Central Khmer 95%</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>Cameroon</td>
<td>English 50%, English 50%</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>Canada</td>
<td>English 58.8%, French 21.6%</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>Cape Verde</td>
<td>Portuguese 100%</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>Central African Republic</td>
<td>Sango 80%, French 20%</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>Chad</td>
<td>Arabic 50%, French 50%</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>Chile</td>
<td>Spanish 100%</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>China</td>
<td>Chinese 100%</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>Colombia</td>
<td>Spanish 100%</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>Congo, Democratic Republic of the</td>
<td>French 33%, Swahili 20%, Lingala 20%</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>Congo, Republic of the</td>
<td>French 30%, Iubi Teke 17%, Lingala 15%</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>Costa Rica</td>
<td>Spanish 100%</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>Cote d’Ivoire</td>
<td>French 50%, Baoulé 14%</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>Croatia</td>
<td>Serbo-Croatian 100%</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>Cuba</td>
<td>Spanish 100%</td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>Cyprus</td>
<td>Greek (Modern) 77%, Turkish 18%</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>Czech Republic</td>
<td>Czech 95.4%, Slovak 1.6%</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>Denmark</td>
<td>Danish 100%</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>Djibouti</td>
<td>Somali 38%, Arabic 20%, French 20%, Afar 13%</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>Dominican Republic</td>
<td>Spanish 100%</td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>Ecuador</td>
<td>Spanish 100%</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>Egypt</td>
<td>Arabic 100%</td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>El Salvador</td>
<td>Spanish 100%</td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>Equatorial Guinea</td>
<td>Spanish 67.6%, French 20%</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>Eritrea</td>
<td>Tigrinya 55%, Tigr 16%</td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>Estonia</td>
<td>Estonian 67.3%, Russian 29.7%, Oromo 33.8%, Amharic 29.3%, Somali 6.2%, Tigr 5.9%, Sidamo 4%</td>
<td></td>
</tr>
<tr>
<td>57</td>
<td>Ethiopia</td>
<td>English 100%</td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>Faroe Islands</td>
<td>Faroese 100%</td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>Fiji</td>
<td>Fiji Hindi 45.3%, Fijian 39.3%</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>Finland</td>
<td>Finnish 91.2%, Swedish 5.5%</td>
<td></td>
</tr>
<tr>
<td>61</td>
<td>France</td>
<td>French 100%</td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>French Guiana</td>
<td>French 100%</td>
<td></td>
</tr>
<tr>
<td>63</td>
<td>Gabon</td>
<td>French 75%, Fang 25%</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>Gambia, The</td>
<td>English 100%</td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>Georgia</td>
<td>Georgian 71%, Russian 9%, Armenian 7%, Azerbaijani 6%</td>
<td></td>
</tr>
<tr>
<td>66</td>
<td>Germany</td>
<td>German 100%</td>
<td></td>
</tr>
<tr>
<td>67</td>
<td>Ghana</td>
<td>Akan 24.7%, English 21.3%, Ewe 12.7%, Abrewn 4.6%</td>
<td></td>
</tr>
<tr>
<td>68</td>
<td>Gibraltar</td>
<td>English 100%</td>
<td></td>
</tr>
<tr>
<td>69</td>
<td>Greece</td>
<td>Greek (Modern) 99%</td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>Greenland</td>
<td>Danish 100%</td>
<td></td>
</tr>
<tr>
<td>71</td>
<td>Grenada</td>
<td>English 87%, French 2%</td>
<td></td>
</tr>
<tr>
<td>72</td>
<td>Guadeloupe</td>
<td>French 99%</td>
<td></td>
</tr>
<tr>
<td>73</td>
<td>Guam</td>
<td>English 38.3%, Chamorro 22.2%, Filipino 22.2%</td>
<td></td>
</tr>
<tr>
<td>74</td>
<td>Guatemala</td>
<td>Spanish 60%</td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>Guinea</td>
<td>French 100%</td>
<td></td>
</tr>
<tr>
<td>76</td>
<td>Guinea-Bissau</td>
<td>Upper Guinea Crioulo 44%, Portuguese 14%</td>
<td></td>
</tr>
<tr>
<td>77</td>
<td>Guyana</td>
<td>English 50%</td>
<td></td>
</tr>
<tr>
<td>78</td>
<td>Haiti</td>
<td>Haitian 75%, French 25%</td>
<td></td>
</tr>
<tr>
<td>79</td>
<td>Honduras</td>
<td>Spanish 100%</td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>Hong Kong</td>
<td>Chinese 95%, English 3.5%</td>
<td></td>
</tr>
<tr>
<td>81</td>
<td>Hungary</td>
<td>Hungarian 93.6%</td>
<td></td>
</tr>
<tr>
<td>82</td>
<td>Iceland</td>
<td>Icelandic 100%</td>
<td></td>
</tr>
<tr>
<td>83</td>
<td>India</td>
<td>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%, Punjabi 2.8%</td>
<td></td>
</tr>
</tbody>
</table>

Table S6 Language demographics by country. Values for each country add to 100% or less (continued next page)
<table>
<thead>
<tr>
<th>Country</th>
<th>Language</th>
<th>Dominant Language</th>
<th>Other Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kuwait</td>
<td>Arabic 100%</td>
<td>Arabic</td>
<td>Arabic</td>
</tr>
<tr>
<td>Kyrgyzstan</td>
<td>Kirghiz 64.7%, Uzbek 13.6%, Russian 12.5%</td>
<td>Kirghiz 25.3%, Uzbek 7.5%, Russian 5.1%</td>
<td>Kirghiz 7.5%, Uzbek 7.5%, Russian 5.1%</td>
</tr>
<tr>
<td>Laos</td>
<td>Lao 100%</td>
<td>Lao</td>
<td>Lao</td>
</tr>
<tr>
<td>Latvia</td>
<td>Latvian 58.2%, Russian 37.5%</td>
<td>Latvian 7.5%, Russian 5.1%</td>
<td>Latvian 7.5%, Russian 5.1%</td>
</tr>
<tr>
<td>Lebanon</td>
<td>Arabic 90%, French 20%</td>
<td>Arabic</td>
<td>Arabic</td>
</tr>
<tr>
<td>Lesotho</td>
<td>Southern Sotho 100%</td>
<td>Sotho</td>
<td>Sotho</td>
</tr>
<tr>
<td>Libya</td>
<td>Arabic 95%</td>
<td>Arabic</td>
<td>Arabic</td>
</tr>
<tr>
<td>Lithuania</td>
<td>Lithuanian 82%, Polish 5.6%, Russian 4%, English 1%</td>
<td>Lithuanian 7.5%, Polish 5.1%, Russian 4%, English 1%</td>
<td>Lithuanian 7.5%, Polish 5.1%, Russian 4%, English 1%</td>
</tr>
<tr>
<td>Luxembourg</td>
<td>Luxembourgish 7%, French 6%, German 4%, English 1%</td>
<td>Luxembourgish 7.5%, French 6%, German 4%, English 1%</td>
<td>Luxembourgish 7.5%, French 6%, German 4%, English 1%</td>
</tr>
<tr>
<td>Macedonia</td>
<td>Macedonian 66.5%, Albanian 25.1%, Turkish 3.5%, Romany 1.9%, Serbo-Croatian 1.2%</td>
<td>Macedonian 66.5%, Albanian 25.1%, Turkish 3.5%, Romany 1.9%, Serbo-Croatian 1.2%</td>
<td>Macedonian 66.5%, Albanian 25.1%, Turkish 3.5%, Romany 1.9%, Serbo-Croatian 1.2%</td>
</tr>
<tr>
<td>Madagascar</td>
<td>French 70%, Malagasy 30%</td>
<td>French</td>
<td>French</td>
</tr>
<tr>
<td>Malawi</td>
<td>Nyamja 70%, Yao 10.1%, Tumbuka 9.5%</td>
<td>Nyamja 70%, Yao 10.1%, Tumbuka 9.5%</td>
<td>Nyamja 70%, Yao 10.1%, Tumbuka 9.5%</td>
</tr>
<tr>
<td>Malaysia</td>
<td>Malay 100%</td>
<td>Malay</td>
<td>Malay</td>
</tr>
<tr>
<td>Maldives</td>
<td>Dhivehi 100%</td>
<td>Dhivehi</td>
<td>Dhivehi</td>
</tr>
<tr>
<td>Mali</td>
<td>Bambara 46.3%, French 10%, Fulah 9.4%, Soninke 6.4%</td>
<td>Bambara 46.3%, French 10%, Fulah 9.4%, Soninke 6.4%</td>
<td>Bambara 46.3%, French 10%, Fulah 9.4%, Soninke 6.4%</td>
</tr>
<tr>
<td>Malta</td>
<td>Maltese 90.2%, English 6%</td>
<td>Maltese</td>
<td>Maltese</td>
</tr>
<tr>
<td>Martinique</td>
<td>French 100%</td>
<td>French</td>
<td>French</td>
</tr>
<tr>
<td>Mauritania</td>
<td>Arabic 100%</td>
<td>Arabic</td>
<td>Arabic</td>
</tr>
<tr>
<td>Mauritius</td>
<td>Bhojpuri 12.1%, French 3.4%, English 1%</td>
<td>Bhojpuri 12.1%, French 3.4%, English 1%</td>
<td>Bhojpuri 12.1%, French 3.4%, English 1%</td>
</tr>
<tr>
<td>Mexico</td>
<td>Spanish 98.5%</td>
<td>Spanish</td>
<td>Spanish</td>
</tr>
<tr>
<td>Moldova</td>
<td>Romanian 76.5%, Russian 11.2%, Ukrainian 4.4%, Gagauz 4%, Bulgarian 1.6%</td>
<td>Romanian 76.5%, Russian 11.2%, Ukrainian 4.4%, Gagauz 4%, Bulgarian 1.6%</td>
<td>Romanian 76.5%, Russian 11.2%, Ukrainian 4.4%, Gagauz 4%, Bulgarian 1.6%</td>
</tr>
<tr>
<td>Monaco</td>
<td>French 100%</td>
<td>French</td>
<td>French</td>
</tr>
<tr>
<td>Mongolia</td>
<td>Mongolian 90%</td>
<td>Mongolian</td>
<td>Mongolian</td>
</tr>
<tr>
<td>Montenegro</td>
<td>Serbo-Croatian 91.1%, Albanian 5.3%</td>
<td>Serbo-Croatian 91.1%, Albanian 5.3%</td>
<td>Serbo-Croatian 91.1%, Albanian 5.3%</td>
</tr>
</tbody>
</table>

**Note:** The table continues with similar entries for other countries and their official languages.
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 (22, 23). 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 (24). 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 (25). The pageid and the Wikipedia API (26) 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. We then validated the places of birth for all people on the list and converted them to a standardized format (e.g., entries such as “NYC”, “New York” or “New York City” were all converted to “New York, NY, US”). 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, and was within a reasonable size that allowed a comprehensive curation and normalization effort. 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 (27) 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.

The following tables show the number of famous people in the Wikipedia 26 dataset for each country (Table S7) and language (Table S8).
<table>
<thead>
<tr>
<th>Country</th>
<th>People (all years)</th>
<th>People (1800-1950)</th>
<th>Country</th>
<th>People (all years)</th>
<th>People (1800-1950)</th>
<th>Country</th>
<th>People (all years)</th>
<th>People (1800-1950)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afghanistan</td>
<td>23</td>
<td>10</td>
<td>Greece</td>
<td>140</td>
<td>34</td>
<td>Nigeria</td>
<td>23</td>
<td>6</td>
</tr>
<tr>
<td>Albania</td>
<td>15</td>
<td>7</td>
<td>Greenland</td>
<td>1</td>
<td></td>
<td>Norway</td>
<td>59</td>
<td>33</td>
</tr>
<tr>
<td>Algeria</td>
<td>17</td>
<td>11</td>
<td>Guadeloupe</td>
<td>4</td>
<td>1</td>
<td>Oman</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Andorra</td>
<td>1</td>
<td></td>
<td>Guam</td>
<td>1</td>
<td>1</td>
<td>Pakistan</td>
<td>28</td>
<td>13</td>
</tr>
<tr>
<td>Angola</td>
<td>5</td>
<td>4</td>
<td>Guatemala</td>
<td>5</td>
<td>2</td>
<td>Palestinian State</td>
<td>14</td>
<td>2</td>
</tr>
<tr>
<td>Antigua and Barbuda</td>
<td>1</td>
<td>1</td>
<td>Guinea</td>
<td>5</td>
<td>3</td>
<td>Panama</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Argentina</td>
<td>102</td>
<td>33</td>
<td>Guinea-Bissau</td>
<td>3</td>
<td>3</td>
<td>Paraguay</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>Armenia</td>
<td>12</td>
<td>4</td>
<td>Guyana</td>
<td>1</td>
<td></td>
<td>Peru</td>
<td>21</td>
<td>12</td>
</tr>
<tr>
<td>Aruba</td>
<td>1</td>
<td>1</td>
<td>Haiti</td>
<td>7</td>
<td>2</td>
<td>Philippines</td>
<td>19</td>
<td>16</td>
</tr>
<tr>
<td>Australia</td>
<td>95</td>
<td>28</td>
<td>Honduras</td>
<td>4</td>
<td>1</td>
<td>Poland</td>
<td>167</td>
<td>114</td>
</tr>
<tr>
<td>Austria</td>
<td>139</td>
<td>91</td>
<td>Hong Kong</td>
<td>5</td>
<td></td>
<td>Portugal</td>
<td>88</td>
<td>16</td>
</tr>
<tr>
<td>Azerbaijan</td>
<td>15</td>
<td>6</td>
<td>Hungary</td>
<td>81</td>
<td>58</td>
<td>Puerto Rico</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Bahrain</td>
<td>1</td>
<td>1</td>
<td>Iceland</td>
<td>15</td>
<td>8</td>
<td>Qatar</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Bangladesh</td>
<td>8</td>
<td>7</td>
<td>India</td>
<td>136</td>
<td>69</td>
<td>Romania</td>
<td>50</td>
<td>26</td>
</tr>
<tr>
<td>Barbados</td>
<td>1</td>
<td></td>
<td>Indonesia</td>
<td>8</td>
<td>7</td>
<td>Russia</td>
<td>369</td>
<td>240</td>
</tr>
<tr>
<td>Belarus</td>
<td>22</td>
<td>10</td>
<td>Iran</td>
<td>61</td>
<td>20</td>
<td>Rwanda</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Belgium</td>
<td>103</td>
<td>40</td>
<td>Iraq</td>
<td>29</td>
<td>8</td>
<td>Saint Kitts and Nevis</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Benin</td>
<td>3</td>
<td>1</td>
<td>Ireland</td>
<td>73</td>
<td>29</td>
<td>Saint Lucia</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Bermuda</td>
<td>1</td>
<td></td>
<td>Isle of Man</td>
<td>4</td>
<td>3</td>
<td>Samoa</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Bhutan</td>
<td>4</td>
<td>1</td>
<td>Israel</td>
<td>73</td>
<td>20</td>
<td>Sao Tome and Principe</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Bolivia</td>
<td>2</td>
<td>1</td>
<td>Latvia</td>
<td>18</td>
<td>11</td>
<td>Saudi Arabia</td>
<td>35</td>
<td>9</td>
</tr>
<tr>
<td>Bosnia and Herzegovina</td>
<td>26</td>
<td>8</td>
<td>Jamaica</td>
<td>10</td>
<td>3</td>
<td>Senegal</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>Botswana</td>
<td>4</td>
<td>3</td>
<td>Japan</td>
<td>137</td>
<td>75</td>
<td>Serbia</td>
<td>60</td>
<td>12</td>
</tr>
<tr>
<td>Brazil</td>
<td>137</td>
<td>53</td>
<td>Jersey</td>
<td>1</td>
<td></td>
<td>Seychelles</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Brunei</td>
<td>1</td>
<td>1</td>
<td>Jordan</td>
<td>7</td>
<td>3</td>
<td>Sierra Leone</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Bulgaria</td>
<td>29</td>
<td>8</td>
<td>Kazakhstan</td>
<td>10</td>
<td>6</td>
<td>Singapore</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>Burkina Faso</td>
<td>3</td>
<td>1</td>
<td>Kenya</td>
<td>10</td>
<td>8</td>
<td>Slovakia</td>
<td>24</td>
<td>6</td>
</tr>
<tr>
<td>Burundi</td>
<td>1</td>
<td></td>
<td>Korea, North</td>
<td>6</td>
<td>4</td>
<td>Slovenia</td>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td>Cambodia</td>
<td>5</td>
<td>2</td>
<td>Korea, South</td>
<td>37</td>
<td>17</td>
<td>Solomon Islands</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Cameroon</td>
<td>10</td>
<td>6</td>
<td>Kosovo</td>
<td>7</td>
<td></td>
<td>Somalia</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>Canada</td>
<td>106</td>
<td>46</td>
<td>Kuwait</td>
<td>3</td>
<td>2</td>
<td>South Africa</td>
<td>45</td>
<td>22</td>
</tr>
<tr>
<td>Cape Verde</td>
<td>4</td>
<td>1</td>
<td>Kyrgyzstan</td>
<td>5</td>
<td>4</td>
<td>South Sudan</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Central African Republic</td>
<td>1</td>
<td>1</td>
<td>Laos</td>
<td>1</td>
<td>1</td>
<td>Spain</td>
<td>298</td>
<td>77</td>
</tr>
<tr>
<td>Chad</td>
<td>2</td>
<td></td>
<td>Latvia</td>
<td>18</td>
<td>11</td>
<td>Sri Lanka</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Chile</td>
<td>27</td>
<td>13</td>
<td>Lebanon</td>
<td>13</td>
<td>6</td>
<td>Sudan</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>China</td>
<td>94</td>
<td>37</td>
<td>Lesotho</td>
<td>1</td>
<td></td>
<td>Suriname</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Colombia</td>
<td>17</td>
<td>3</td>
<td>Liberia</td>
<td>5</td>
<td>2</td>
<td>Swaziland</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Congo, Democratic Republic of the</td>
<td>7</td>
<td>3</td>
<td>Libya</td>
<td>11</td>
<td>2</td>
<td>Sweden</td>
<td>135</td>
<td>61</td>
</tr>
<tr>
<td>Congo, Republic of</td>
<td>2</td>
<td></td>
<td>Lithuania</td>
<td>28</td>
<td>19</td>
<td>Switzerland</td>
<td>102</td>
<td>56</td>
</tr>
<tr>
<td>Costa Rica</td>
<td>3</td>
<td>1</td>
<td>Luxembourg</td>
<td>8</td>
<td>4</td>
<td>Syria</td>
<td>19</td>
<td>2</td>
</tr>
<tr>
<td>Cote d'Ivoire</td>
<td>15</td>
<td>3</td>
<td>Macedonia</td>
<td>15</td>
<td>3</td>
<td>Taiwan</td>
<td>19</td>
<td>4</td>
</tr>
<tr>
<td>Croatia</td>
<td>56</td>
<td>10</td>
<td>Madagascar</td>
<td>2</td>
<td>1</td>
<td>Tajikistan</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Cuba</td>
<td>13</td>
<td>9</td>
<td>Malawi</td>
<td>4</td>
<td>4</td>
<td>Tanzania</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Cyprus</td>
<td>9</td>
<td>3</td>
<td>Malaysia</td>
<td>6</td>
<td>4</td>
<td>Thailand</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>105</td>
<td>53</td>
<td>Maldives</td>
<td>3</td>
<td>1</td>
<td>Turks and Caicos Islands</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Denmark</td>
<td>99</td>
<td>39</td>
<td>Mali</td>
<td>8</td>
<td>4</td>
<td>Togo</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Djibouti</td>
<td>1</td>
<td></td>
<td>Malta</td>
<td>3</td>
<td>2</td>
<td>Tonga</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Dominican Republic</td>
<td>2</td>
<td>1</td>
<td>Martinique</td>
<td>3</td>
<td>2</td>
<td>Trinidad and Tobago</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Ecuador</td>
<td>4</td>
<td>1</td>
<td>Mauritania</td>
<td>1</td>
<td></td>
<td>Tunisia</td>
<td>18</td>
<td>7</td>
</tr>
<tr>
<td>Egypt</td>
<td>68</td>
<td>24</td>
<td>Mauritius</td>
<td>1</td>
<td>1</td>
<td>Turkey</td>
<td>184</td>
<td>35</td>
</tr>
<tr>
<td>El Salvador</td>
<td>3</td>
<td>1</td>
<td>Mexico</td>
<td>56</td>
<td>23</td>
<td>Turkmenistan</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Equatorial Guinea</td>
<td>1</td>
<td>1</td>
<td>Micronesia,</td>
<td>1</td>
<td>1</td>
<td>Uganda</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Eritrea</td>
<td>1</td>
<td></td>
<td>Federated States</td>
<td>1</td>
<td></td>
<td>Ukraine</td>
<td>100</td>
<td>58</td>
</tr>
<tr>
<td>Estonia</td>
<td>12</td>
<td>9</td>
<td>Moldova</td>
<td>5</td>
<td>2</td>
<td>United Arab Emirates</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Ethiopia</td>
<td>10</td>
<td>6</td>
<td>Monaco</td>
<td>4</td>
<td>1</td>
<td>United Kingdom</td>
<td>1,140</td>
<td>508</td>
</tr>
<tr>
<td>Faroe Islands</td>
<td>4</td>
<td>1</td>
<td>Mongolia</td>
<td>8</td>
<td>1</td>
<td>United States</td>
<td>2,291</td>
<td>1,221</td>
</tr>
<tr>
<td>Finland</td>
<td>63</td>
<td>34</td>
<td>Morocco</td>
<td>14</td>
<td>7</td>
<td>Uruguay</td>
<td>23</td>
<td>7</td>
</tr>
<tr>
<td>France</td>
<td>857</td>
<td>397</td>
<td>Mozambique</td>
<td>6</td>
<td>3</td>
<td>Uzbekistan</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>French Guiana</td>
<td>1</td>
<td></td>
<td>Namibia</td>
<td>2</td>
<td>2</td>
<td>Vanuatu</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Gabon</td>
<td>3</td>
<td>3</td>
<td>Nauru</td>
<td>1</td>
<td></td>
<td>Venezuela</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>Gambia, The</td>
<td>1</td>
<td></td>
<td>Nepal</td>
<td>4</td>
<td>3</td>
<td>Vietnam</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>Georgia</td>
<td>21</td>
<td>12</td>
<td>Netherlands</td>
<td>162</td>
<td>56</td>
<td>Virgin Islands</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Germany</td>
<td>740</td>
<td>407</td>
<td>New Caledonia</td>
<td>2</td>
<td></td>
<td>Yemen</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Ghana</td>
<td>10</td>
<td>4</td>
<td>New Zealand</td>
<td>17</td>
<td>9</td>
<td>Zambia</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Gibraltar</td>
<td>1</td>
<td></td>
<td>Nicaragua</td>
<td>5</td>
<td>5</td>
<td>Zimbabwe</td>
<td>7</td>
<td>4</td>
</tr>
</tbody>
</table>

Table S7 Number of people with articles in at least 26 Wikipedia language editions, by country.
<table>
<thead>
<tr>
<th>Language</th>
<th>Code</th>
<th>People (all years)</th>
<th>People (1800-1950)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Afrikaans</td>
<td>afr</td>
<td>6.94</td>
<td>4.14</td>
</tr>
<tr>
<td>2 Albanian</td>
<td>sqi</td>
<td>26.87</td>
<td>8.34</td>
</tr>
<tr>
<td>3 Arabic</td>
<td>ara</td>
<td>273.07</td>
<td>94.46</td>
</tr>
<tr>
<td>4 Armenian</td>
<td>aze</td>
<td>13.42</td>
<td>4.84</td>
</tr>
<tr>
<td>5 Azerbaijani</td>
<td>bye</td>
<td>25.79</td>
<td>9.74</td>
</tr>
<tr>
<td>6 Basque</td>
<td>eus</td>
<td>5.96</td>
<td>1.54</td>
</tr>
<tr>
<td>7 Belarusian</td>
<td>bel</td>
<td>5.15</td>
<td>2.34</td>
</tr>
<tr>
<td>8 Bengali</td>
<td>ben</td>
<td>18.86</td>
<td>12.45</td>
</tr>
<tr>
<td>9 Bulgarian</td>
<td>bul</td>
<td>22.35</td>
<td>6.18</td>
</tr>
<tr>
<td>10 Catalan</td>
<td>cat</td>
<td>51.06</td>
<td>13.09</td>
</tr>
<tr>
<td>11 Chinese</td>
<td>zho</td>
<td>115.6</td>
<td>44.24</td>
</tr>
<tr>
<td>12 Czech</td>
<td>ces</td>
<td>100.17</td>
<td>50.56</td>
</tr>
<tr>
<td>13 Danish</td>
<td>dan</td>
<td>100</td>
<td>39</td>
</tr>
<tr>
<td>14 Dutch</td>
<td>nld</td>
<td>226.86</td>
<td>81.26</td>
</tr>
<tr>
<td>15 English</td>
<td>eng</td>
<td>3300.8</td>
<td>1617.77</td>
</tr>
<tr>
<td>16 Estonian</td>
<td>est</td>
<td>10.1</td>
<td>6.06</td>
</tr>
<tr>
<td>17 Filipino</td>
<td>fil</td>
<td>19.22</td>
<td>16.22</td>
</tr>
<tr>
<td>18 Finnish</td>
<td>fin</td>
<td>57.46</td>
<td>31.01</td>
</tr>
<tr>
<td>19 French</td>
<td>fra</td>
<td>997.7</td>
<td>455.51</td>
</tr>
<tr>
<td>20 Galician</td>
<td>glg</td>
<td>20.86</td>
<td>5.39</td>
</tr>
<tr>
<td>21 Georgian</td>
<td>kat</td>
<td>14.91</td>
<td>8.52</td>
</tr>
<tr>
<td>22 German</td>
<td>deu</td>
<td>929.09</td>
<td>524.1</td>
</tr>
<tr>
<td>23 Haitian</td>
<td>hat</td>
<td>5.25</td>
<td>1.5</td>
</tr>
<tr>
<td>24 Hebrew</td>
<td>heb</td>
<td>58.4</td>
<td>16</td>
</tr>
<tr>
<td>25 Hindi</td>
<td>hin</td>
<td>55.95</td>
<td>28.39</td>
</tr>
<tr>
<td>26 Hungarian</td>
<td>hun</td>
<td>84.01</td>
<td>57.13</td>
</tr>
<tr>
<td>27 Icelandic</td>
<td>isl</td>
<td>15</td>
<td>8</td>
</tr>
<tr>
<td>28 Italian</td>
<td>ita</td>
<td>801.15</td>
<td>198.09</td>
</tr>
<tr>
<td>29 Japanese</td>
<td>jpn</td>
<td>137</td>
<td>75</td>
</tr>
<tr>
<td>30 Kazakh</td>
<td>kaz</td>
<td>6.3</td>
<td>3.78</td>
</tr>
<tr>
<td>31 Kirghiz</td>
<td>kir</td>
<td>3.23</td>
<td>2.59</td>
</tr>
<tr>
<td>32 Korean</td>
<td>kor</td>
<td>43</td>
<td>21</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language</th>
<th>Code</th>
<th>People (all years)</th>
<th>People (1800-1950)</th>
</tr>
</thead>
<tbody>
<tr>
<td>33 Latvian</td>
<td>lav</td>
<td>10.48</td>
<td>6.4</td>
</tr>
<tr>
<td>34 Lithuanian</td>
<td>lit</td>
<td>22.96</td>
<td>15.58</td>
</tr>
<tr>
<td>35 Macedonian</td>
<td>mkd</td>
<td>9.97</td>
<td>2</td>
</tr>
<tr>
<td>36 Malay</td>
<td>msa</td>
<td>15.99</td>
<td>12.56</td>
</tr>
<tr>
<td>37 Malayalam</td>
<td>mal</td>
<td>4.35</td>
<td>2.21</td>
</tr>
<tr>
<td>38 Maltese</td>
<td>mlt</td>
<td>2.71</td>
<td>1.8</td>
</tr>
<tr>
<td>39 Maori</td>
<td>mri</td>
<td>0.66</td>
<td>0.35</td>
</tr>
<tr>
<td>40 Marathi</td>
<td>mar</td>
<td>9.52</td>
<td>4.83</td>
</tr>
<tr>
<td>41 Modern Greek</td>
<td>ell</td>
<td>147.22</td>
<td>38.08</td>
</tr>
<tr>
<td>42 Mongolian</td>
<td>mon</td>
<td>7.2</td>
<td>0.9</td>
</tr>
<tr>
<td>43 Norwegian</td>
<td>nor</td>
<td>59</td>
<td>33</td>
</tr>
<tr>
<td>44 Persian</td>
<td>pers</td>
<td>42.83</td>
<td>15.6</td>
</tr>
<tr>
<td>45 Polish</td>
<td>pol</td>
<td>164.89</td>
<td>112.56</td>
</tr>
<tr>
<td>46 Portuguese</td>
<td>por</td>
<td>235.69</td>
<td>74.92</td>
</tr>
<tr>
<td>47 Romanian</td>
<td>ron</td>
<td>49.33</td>
<td>25.19</td>
</tr>
<tr>
<td>48 Russian</td>
<td>russ</td>
<td>429.38</td>
<td>272.91</td>
</tr>
<tr>
<td>49 Serbo-Croatian</td>
<td>hbs</td>
<td>152.84</td>
<td>36.92</td>
</tr>
<tr>
<td>50 Sinhala</td>
<td>sin</td>
<td>4.44</td>
<td>3.7</td>
</tr>
<tr>
<td>51 Slovak</td>
<td>slovak</td>
<td>21.82</td>
<td>5.39</td>
</tr>
<tr>
<td>52 Slovenian</td>
<td>slv</td>
<td>13.66</td>
<td>2.73</td>
</tr>
<tr>
<td>53 Spanish</td>
<td>spa</td>
<td>774.64</td>
<td>305.48</td>
</tr>
<tr>
<td>54 Swahili</td>
<td>swa</td>
<td>12.4</td>
<td>10</td>
</tr>
<tr>
<td>55 Swedish</td>
<td>swe</td>
<td>138.47</td>
<td>62.87</td>
</tr>
<tr>
<td>56 Tajik</td>
<td>tajik</td>
<td>14.91</td>
<td>8.52</td>
</tr>
<tr>
<td>57 Tamil</td>
<td>tam</td>
<td>9.33</td>
<td>5.1</td>
</tr>
<tr>
<td>58 Thai</td>
<td>th</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>59 Turkish</td>
<td>tur</td>
<td>164.86</td>
<td>33.64</td>
</tr>
<tr>
<td>60 Turkmen</td>
<td>tuk</td>
<td>3.21</td>
<td>1.22</td>
</tr>
<tr>
<td>61 Ukrainian</td>
<td>ukr</td>
<td>67.46</td>
<td>39.01</td>
</tr>
<tr>
<td>62 Urdu</td>
<td>urd</td>
<td>9.04</td>
<td>4.49</td>
</tr>
<tr>
<td>63 Uzbek</td>
<td>uzb</td>
<td>8.9</td>
<td>1.98</td>
</tr>
<tr>
<td>64 Vietnamese</td>
<td>vie</td>
<td>10.95</td>
<td>9.28</td>
</tr>
</tbody>
</table>

Table S8 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* (21) 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 famous people for each language, 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.

The following tables show the number of famous people in the *Human Accomplishment* dataset for each country (Table S9) and language (Table S10).
<table>
<thead>
<tr>
<th>Country</th>
<th>People (all years)</th>
<th>People (1800-1950)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  Ancient Greece</td>
<td>134</td>
<td>N/A</td>
</tr>
<tr>
<td>2  Arab World</td>
<td>86</td>
<td>14</td>
</tr>
<tr>
<td>3  Argentina</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4  Australia</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5  Austria</td>
<td>75</td>
<td>48</td>
</tr>
<tr>
<td>6  Belgium</td>
<td>82</td>
<td>27</td>
</tr>
<tr>
<td>7  Brazil</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>8  Bulgaria</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9  Canada</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>10 Colombia</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>11 China</td>
<td>237</td>
<td>22</td>
</tr>
<tr>
<td>12 Croatia</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>13 Cuba</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>14 Czech Republic</td>
<td>48</td>
<td>28</td>
</tr>
<tr>
<td>15 Denmark</td>
<td>37</td>
<td>20</td>
</tr>
<tr>
<td>16 Finland</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>17 France</td>
<td>542</td>
<td>236</td>
</tr>
<tr>
<td>18 Germany</td>
<td>536</td>
<td>267</td>
</tr>
<tr>
<td>19 Greece</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>20 Guatemala</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>21 Hungary</td>
<td>21</td>
<td>18</td>
</tr>
<tr>
<td>22 Iceland</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>23 India</td>
<td>93</td>
<td>16</td>
</tr>
<tr>
<td>24 Italy</td>
<td>389</td>
<td>58</td>
</tr>
</tbody>
</table>

Table S9 Number of people listed on human accomplishment, by country.

<table>
<thead>
<tr>
<th>Language</th>
<th>Code</th>
<th>People (all years)</th>
<th>People (1800-1950)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Afrikaans</td>
<td>afr</td>
<td>0.13</td>
<td>0.13</td>
</tr>
<tr>
<td>2 Albanian</td>
<td>sqi</td>
<td>0.88</td>
<td>0.47</td>
</tr>
<tr>
<td>3 Arabic</td>
<td>ara</td>
<td>86.05</td>
<td>14.05</td>
</tr>
<tr>
<td>4 Basque</td>
<td>eus</td>
<td>1.52</td>
<td>0.52</td>
</tr>
<tr>
<td>5 Bengali</td>
<td>ben</td>
<td>7.53</td>
<td>1.3</td>
</tr>
<tr>
<td>6 Bulgarian</td>
<td>bul</td>
<td>0.77</td>
<td>0.77</td>
</tr>
<tr>
<td>7 Catalan</td>
<td>cat</td>
<td>12.92</td>
<td>4.42</td>
</tr>
<tr>
<td>8 Chinese</td>
<td>zho</td>
<td>237.16</td>
<td>22.16</td>
</tr>
<tr>
<td>9 Czech</td>
<td>ces</td>
<td>45.79</td>
<td>26.71</td>
</tr>
<tr>
<td>10 Dutch</td>
<td>ndl</td>
<td>133.2</td>
<td>47.2</td>
</tr>
<tr>
<td>11 English</td>
<td>eng</td>
<td>788.1</td>
<td>466.26</td>
</tr>
<tr>
<td>12 Finnish</td>
<td>fin</td>
<td>5.47</td>
<td>4.56</td>
</tr>
<tr>
<td>13 French</td>
<td>fra</td>
<td>590.27</td>
<td>255.74</td>
</tr>
<tr>
<td>14 Galician</td>
<td>glg</td>
<td>5.32</td>
<td>1.82</td>
</tr>
<tr>
<td>15 German</td>
<td>deu</td>
<td>643.22</td>
<td>329.91</td>
</tr>
<tr>
<td>16 Greek</td>
<td>grc</td>
<td>134</td>
<td></td>
</tr>
<tr>
<td>17 Greek (Modern)</td>
<td>ell</td>
<td>8.96</td>
<td>5.99</td>
</tr>
<tr>
<td>18 Hindi</td>
<td>hin</td>
<td>38.16</td>
<td>6.59</td>
</tr>
<tr>
<td>19 Hungarian</td>
<td>hun</td>
<td>20.5</td>
<td>17.62</td>
</tr>
<tr>
<td>20 Icelandic</td>
<td>isl</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>21 Italian</td>
<td>ita</td>
<td>393.22</td>
<td>60.14</td>
</tr>
</tbody>
</table>

Table S10 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 S4 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.
Figure S4 Correlation of the Wikipedia 26 and Human Accomplishment datasets
References for the SI Appendix


26. Wikimedia MediaWiki API. Available at: https://www.mediawiki.org/wiki/API.