Online Appendix

A Additional Tables

Table A.1: Linguistic Distance from English and GDP

	Log GDP per Capita					
	(1) 1870	(2) 1913	(3) 2018	(4) 1870	(5) 1913	(6) 2018
Log Physical Distance between Country and the UK	-0.170*** (0.058)	-0.207*** (0.064)	-0.237*** (0.066)	-0.248*** (0.054)	-0.315*** (0.065)	-0.323*** (0.072)
Number of Weeks Required to Learn the Plurality Language	-0.010*** (0.002)	-0.013*** (0.003)	-0.008* (0.004)	-0.005** (0.002)	-0.007*** (0.003)	-0.003 (0.005)
Observations	61	61	61	55	55	55
R^2	0.395	0.428	0.208	0.369	0.426	0.198
Includes English-speaking Countries	✓	✓	✓			

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: GDP per capita is from the Maddison Project. The physical distance between the region and the UK is from *CEPII* database using the Great Circle Formula. The number of weeks an English native speaker will take to obtain "Professional Working Proficiency" in the plurality language of a country is estimated by the U.S. Department of State's Foreign Service Institute. See Appendix D for data construction and sources. Robust standard errors are in parentheses. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

Table A.2: Annualized Export/Productivity Growth and British Patent Relevance - British Colonies and Steam Intensity

	Export Growth		Γ_{ik}	
	(1)	(2)	(3)	(4)
BPR × Japan	3.027^{***}	3.175^{**}	0.281^{***}	0.194^{*}
	(0.791)	(1.242)	(0.087)	(0.106)
BPR × Not Japan	-0.871^{***}	-0.155	-0.026	-0.001
	(0.242)	(0.259)	(0.028)	(0.028)
BPR × British Colony	0.694		0.077	
	(0.488)		(0.054)	
Steam Intensity		-0.736^{**}		0.015
		(0.296)		(0.035)
Observations	1395	690	1244	627
R^2	0.234	0.310	0.011	0.067
Country fixed effects	✓	✓	✓	✓
Sample	All	All	All	All

Note: The dependent variable, "Export Growth," is the annualized export growth rate for industry k in region i between {1880,1885} and {1905,1910}. Γ_{ik}, is the annualized productivity growth rate for industry k in region i between {1880,1885} and {1905,1910}. "British Patent Relevance" is a variable that captures how relevant the titles of British patents (1617-1852) are to the vocabulary of an industry k. The Japan dummy equals one if the region is Japan and zero otherwise, "Not Japan" is analogously defined. "British Colony" is a dummy for whether a region was a British colony in the 1880-1910 window. Steam Intensity is constructed as Steam Engine Horsepower/Wage Bill by industry using French manufacturing census data from the 1860s (see Appendix D.9 for details about the data construction). Robust standard errors are in parentheses. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

Table A.3: Annualized Export Growth and British Patent Relevance - Manufacturing Sectors

	Export Growth						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
BPR × Japan	2.775**	2.775**	2.775**	2.775**	2.775**	2.775**	2.775**
	(1.161)	(1.156)	(1.157)	(1.157)	(1.157)	(1.158)	(1.159)
BPR × Not Japan		-0.348	-0.348	-0.544*	-1.004**		
		(0.239)	(0.274)	(0.286)	(0.443)		
BPR × English-Speaking			0.004				
			(0.515)				
BPR × French-Speaking				0.883*			
				(0.487)			
BPR × Top-4 Codified					1.230**		
					(0.501)		
BPR × High-Income						-0.259	-0.259
						(0.261)	(0.261)
BPR × Medium-Income						-0.135	-0.065
						(0.663)	(0.682)
BPR × Low-Income						-0.809	-0.478
						(0.571)	(0.760)
BPR × Asia							-0.731
							(0.872)
Observations	31	661	661	661	661	661	661
R^2	0.160	0.364	0.364	0.366	0.369	0.365	0.365
Country fixed effects	✓	✓	✓	✓	✓	✓	✓
Sample	Japan	All	All	All	All	All	All

Note: The dependent variable, "Export Growth," is the annualized export growth rate for industry k in region i between {1880,1885} and {1905,1910}. "British Patent Relevance" is a variable that captures how relevant the titles of British patents (1617-1852) are to the vocabulary of an industry k. Japan dummy equals one if the region is Japan and zero otherwise, "Not Japan" is analogously defined. "English-speaking" is an indicator equal to 1 if the region's plurality language is English. "Top-4 Codified" is a dummy for countries that speak one of the four most codified languages: French, English, German, and Italian. {High, Medium, Low}Income are indicator variables which use 1870 GDP per capita from the Maddison Project to identify if a region is in the top third of the income distribution (high), middle third (medium), or in the bottom third (bottom); we set these dummies to 0 for Japan. Asia dummy equals 1 if the region is in Asia and 0 if it is Japan or not in Asia. Robust standard errors are in parentheses. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

Table A.4: Annualized Productivity Growth and British Patent Relevance - Manufacturing Sectors

				Γ_{ik}			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
BPR × Japan	0.256**	0.256**	0.256**	0.256**	0.256**	0.256**	0.256**
	(0.114)	(0.112)	(0.112)	(0.112)	(0.112)	(0.112)	(0.113)
BPR × Not Japan		0.000	0.001	-0.008	-0.027		
		(0.026)	(0.029)	(0.031)	(0.047)		
BPR × English-Speaking			-0.003				
			(0.057)				
BPR × French-Speaking				0.037			
				(0.052)			
BPR × Top-4 Codified					0.053		
					(0.053)		
BPR × High-Income						-0.015	-0.015
						(0.027)	(0.027)
BPR × Medium-Income						0.099	0.115*
						(0.067)	(0.069)
BPR × Low-Income						-0.059	0.011
						(0.066)	(0.082)
BPR × Asia							-0.157
							(0.098)
Observations	24	587	587	587	587	587	587
R^2	0.080	0.120	0.120	0.120	0.121	0.126	0.130
Country fixed effects	✓	✓	✓	✓	✓	✓	✓
Sample	Japan	All	All	All	All	All	All

Note: The dependent variable, Γ_{ik}, is the annualized productivity growth rate for industry k in region i between {1880,1885} and {1905,1910}. "British Patent Relevance" is a variable that captures how relevant the titles of British patents (1617-1852) are to the vocabulary of an industry k. Japan dummy equals one if the region is Japan and zero otherwise, "Not Japan" is analogously defined. "English-speaking" is an indicator equal to 1 if the region's plurality language is English. "Top-4 Codified" is a dummy for countries that speak one of the four most codified languages: French, English, German, and Italian. Steam Intensity is constructed as Steam Engine Horsepower/Wage Bill at an industry level. {High, Medium, Low}Income are indicator variables which use 1870 GDP per capita from the Maddison Project to identify if a region is in the top third of the income distribution (high), middle third (medium), or in the bottom third (bottom); we set these dummies to 0 for Japan. Asia dummy equals 1 if the region is in Asia and 0 if it is Japan or not in Asia. Robust standard errors are in parentheses. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

Table A.5: Annualized Export Growth and British Patent Relevance: Selected Countries

	Export Growth							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
British Patent Relevance	0.190	0.323	1.105*	0.572	0.340*	0.418	-4.105	-1.273*
	(0.325)	(0.699)	(0.624)	(0.436)	(0.193)	(0.395)	(2.477)	(0.643)
Observations	86	48	88	72	74	86	29	46
R^2	0.002	0.002	0.024	0.016	0.032	0.005	0.146	0.040
Constant	✓	✓	✓	✓	✓	✓	✓	✓
Sample	France	Belgium	UK	US	Italy	Germany	Spain	Mexico

Note: The dependent variable, "Export Growth," is the annualized export growth rate for industry k in region i between {1880,1885} and {1905,1910}. "British Patent Relevance" is a variable that captures how relevant the titles of British patents (1617-1852) are to the vocabulary of an industry k. Robust standard errors are in parentheses. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

Table A.6: Annualized Productivity Growth and British Patent Relevance: Selected Countries

	Γ_{ik}							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
British Patent Relevance	-0.022	0.068	0.100*	0.019	0.106*	0.032	-0.375	0.013
	(0.034)	(0.064)	(0.052)	(0.083)	(0.060)	(0.046)	(0.252)	(0.082)
Observations	73	43	74	63	62	73	26	42
R^2	0.002	0.008	0.033	0.000	0.024	0.004	0.152	0.001
Constant	✓	✓	✓	✓	✓	✓	✓	✓
Sample	France	Belgium	UK	US	Italy	Germany	Spain	Mexico

Note: The dependent variable, Γ_{ik}, is the annualized productivity growth rate for industry k in region i between {1880,1885} and {1905,1910}. "British Patent Relevance" is a variable that captures how relevant the titles of British patents (1617-1852) are to the vocabulary of an industry k. Robust standard errors are in parentheses. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

Table A.7: Annualized Export Growth and British Patent Relevance: Dropping Regions

	Export Growth, Dropping Exports to						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	English-Speaking	British Colonies	Languages Similar to English	High-Income	Medium-Income	Low-Income	Asian
British Patent Relevance	2.850***	2.850***	2.856***	2.351***	2.869***	2.674***	2.674***
	(0.776)	(0.776)	(0.777)	(0.758)	(0.792)	(0.937)	(0.937)
Observations	71	71	71	70	67	61	61
R^2	0.118	0.118	0.119	0.077	0.121	0.074	0.074
Constant	✓	✓	✓	✓	✓	✓	✓
Sample	Japan	Japan	Japan	Japan	Japan	Japan	Japan

Note: The dependent variable, "Export Growth," is the annualized export growth rate for industry k in region i between {1880,1885} and {1905,1910}. "British Patent Relevance" is a variable that captures how relevant the titles of British patents (1617-1852) are to the vocabulary of an industry k. Each column drops exports to a different subset of countries/regions. (1) Drops English-Speaking countries. (2) Drops British Colonies. (3) Drops countries with a language similar to English, defined as those where it takes six or fewer months for an English speaker to become proficient. (4), (5), and (6) drop High, Medium, and Low-income countries, respectively. (7) Drops exports to Asian countries. Robust standard errors are in parentheses. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

Table A.8: Annualized Export/Productivity Growth and British Patent Relevance: Dropping Sectors

	Export Growth, Dropping			Productivity Growth, Dropping		
	(1)	(2)	(3)	(4)	(5)	(6)
	Cotton-Textiles	All Textiles	Iron and Fabricated Metals	Cotton-Textiles	All Textiles	Iron and Fabricated Metals
British Patent Relevance	3.395^{***}	3.662^{***}	3.135^{***}	0.362^{***}	0.306^{*}	0.289^{***}
	(0.948)	(1.368)	(0.829)	(0.095)	(0.156)	(0.092)
Observations	69	63	69	54	49	55
R^2	0.119	0.091	0.123	0.083	0.037	0.068
Constant	✓	✓	✓	✓	✓	✓
Sample	Japan	Japan	Japan	Japan	Japan	Japan

Note: The dependent variable, "Export Growth," is the annualized export growth rate for industry k in region i between {1880,1885} and {1905,1910}. "British Patent Relevance" (BPR) is a variable that captures how relevant the titles of British patents (1617-1852) are to the vocabulary of an industry k. Each column drops exports to a particular industry or group of industries. (1) and (4) drops cotton textile-related industries. (2) and (5) drops all industries related to textiles. (3) and (6) drops industries related to producing iron. Robust standard errors are in parentheses. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

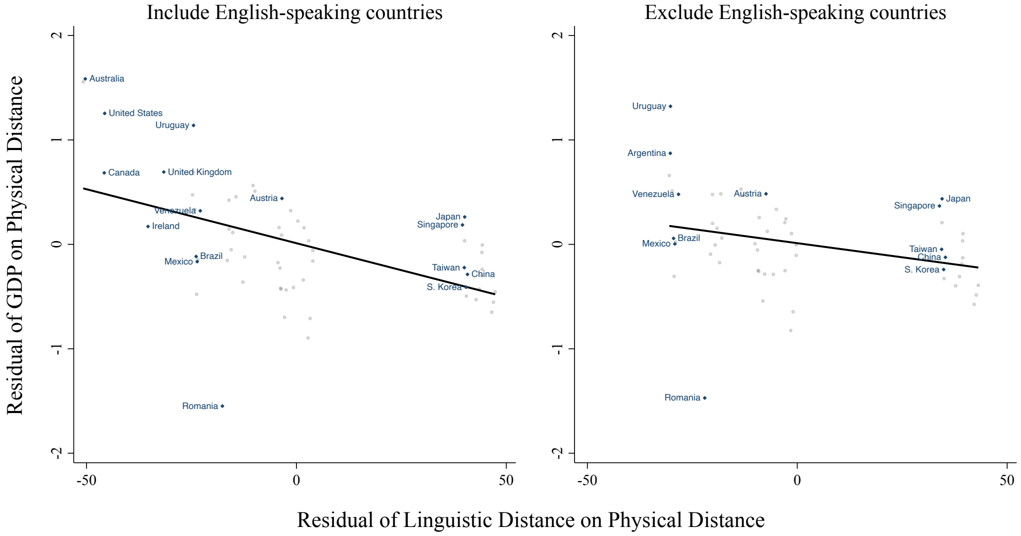
Table A.9: Japanese export growth and British Patent Relevance 1875-1910

	Annualized Export Growth Between 1875 and						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	1880	1885	1890	1895	1900	1905	1910
British Patent Relevance	-3.246^{**}	-0.851	0.168	0.633^{**}	0.493^{**}	0.471^{**}	0.342^{**}
	(1.596)	(0.575)	(0.423)	(0.280)	(0.226)	(0.186)	(0.153)
Observations	40	45	46	47	45	46	47
Constant	✓	✓	✓	✓	✓	✓	✓

Note: The dependent variable is annualized Japanese export growth for the year reported relative to 1875. The number of observations changes across specifications because of the different number of traded sectors in different years. Robust standard errors in parentheses: $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

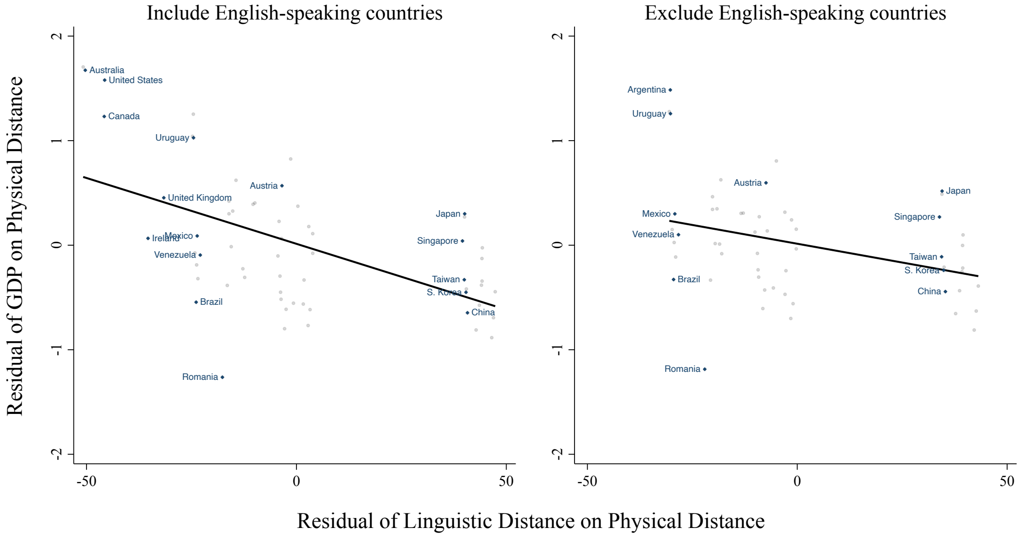
B Additional Figures

Figure A.1: Linguistic Distance Partial Regression Plot for 1870



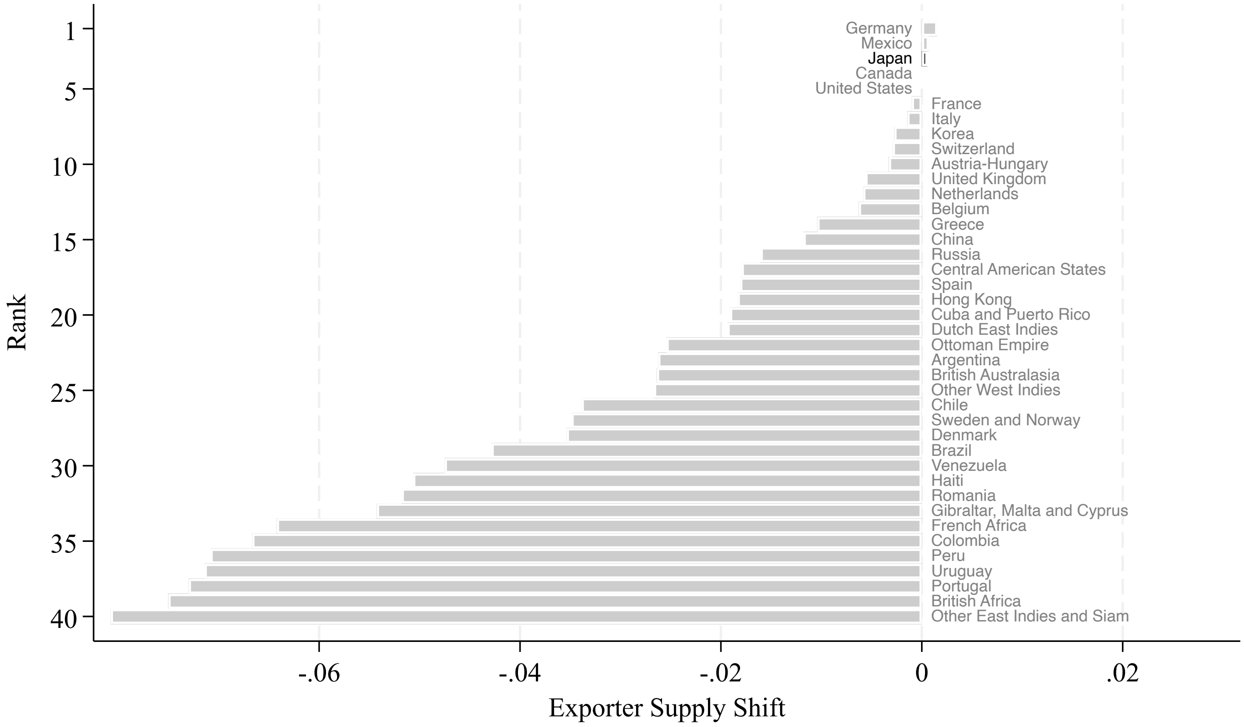
Note: This figure plots the relationship between log GDP per capita in 1870 and linguistic distance after controlling for log physical distance. Data are from the Maddison dataset, the U.S. Department of State's Foreign Service Institute, and *CEPII*, respectively.

Figure A.2: Linguistic Distance Partial Regression Plot for 1913



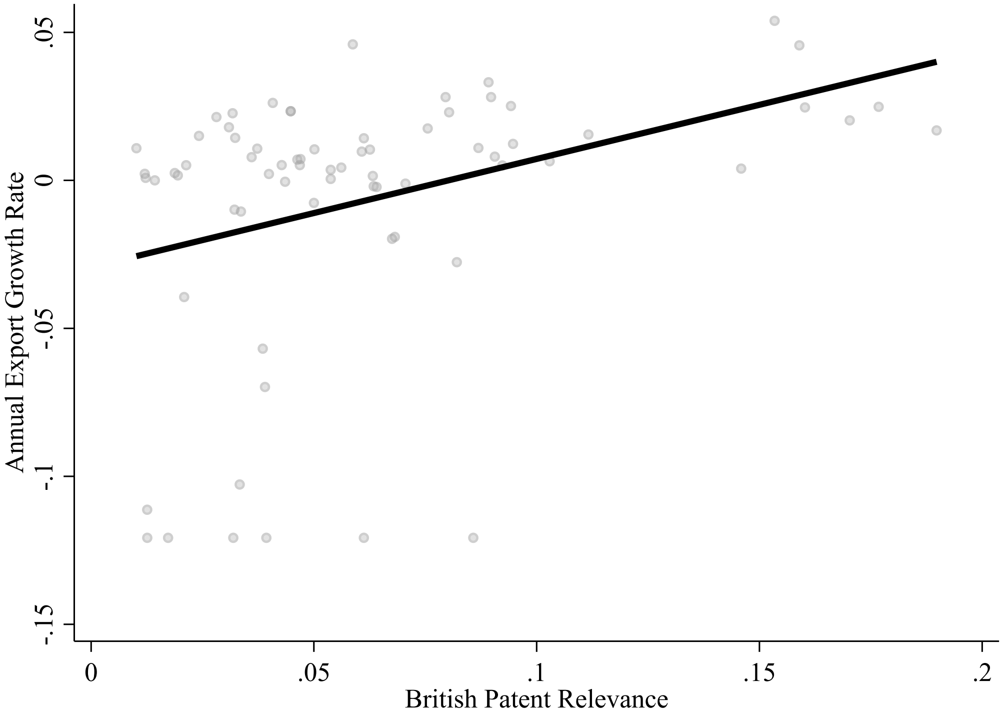
Note: This figure plots the relationship between log GDP per capita in 1913 and linguistic distance after controlling for log physical distance. Data are from the Maddison dataset, the U.S. Department of State's Foreign Service Institute, and *CEPII*, respectively.

Figure A.3: Relative Annualized Exporter Supply Shift by Exporter



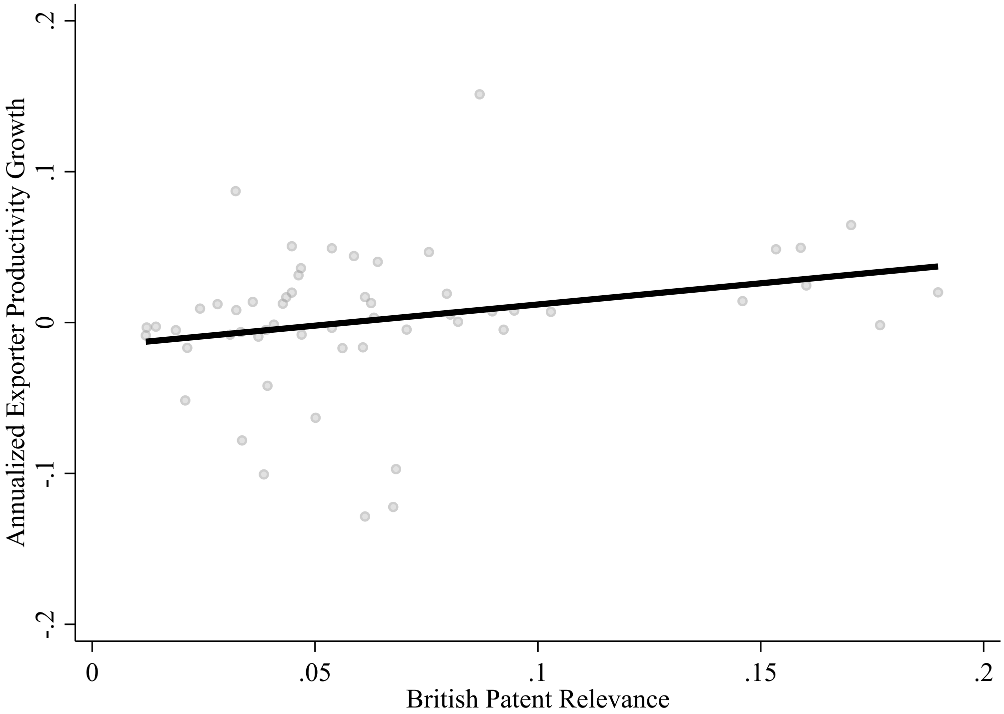
Note: Annualized per-capita exporter supply shifts are expressed as relative to the US, i.e., they are defined as $\hat{\gamma}_i - \hat{\gamma}_{US}$. See text for details on variable construction.

Figure A.4: Annualized Export Growth and British Patent Relevance for Japan



Note: The dependent variable, "Export Growth," is the annualized export growth rate for industry *k* between {1880,1885} and {1905,1910}. British Patent Relevance is a variable that captures how relevant the titles of British patents (1617-1852) are to the vocabulary of an industry *k*. See text for details on variable construction.

Figure A.5: Annualized Prod. Growth Γ and British Patent Relevance for Japan



Note: The dependent variable, Γ_{ik}, is the annualized growth rate in comparative advantage for industry *k* in region *i* between {1880,1885} and {1905,1910}. "British Patent Relevance" is a variable that captures how relevant the titles of British patents (1617-1852) are to the vocabulary of an industry *k*. See text for details on variable construction.

C Constructing Annual Growth Rates

We build the bilateral global trade data by merging bilateral industry export flows from different source countries (Belgium, Japan, Italy, or the U.S.). These data source countries sometimes only report exports in an industry in one of the early years (1880 or 1885) or one of the later years (1905 or 1910). Rather than throw out the industry for all countries when 1880 or 1910 is not reported by one source region, we adopt a procedure to let us be flexible about the start and end dates by computing the average annual export growth rates between any of two potential start years at the beginning of our sample (1880 or 1885) and any of two potential end years at the end of our sample (1905 or 1910).

We set the start year equal to 1880 if the source region reports data in that year or 1885 if data is not available for 1880 but is available for 1885. Similarly, we set the final year equal to 1910 if the source region reports data for that year or 1905 if data is not available for 1910 but is available for 1905. Since this means that the start and final years for bilateral trade growth rates can vary by data source region, we annualize the data so our export and productivity growth rates can be interpreted as average annual growth rates.

We use two procedures to annualize the data. If the reporting region exports the product in 1880 or 1885 (i.e., $\sum_j x_{ijks} > 0$ for $s = 1880$ or 1885), we set s equal to the first year that satisfies $\sum_j x_{ijks} > 0$. We drop the sector if $\sum_j x_{ijks} = 0$ because industry growth rates are undefined if a country does not export anything in the industry in the first period. Similarly, we set f equal to the last year ($f \in \{1905, 1910\}$) that satisfies $\sum_j x_{ijkf} > 0$. We compute the annual growth rate for all bilateral exports satisfying $x_{ijks} > 0$ as

$$g^C_{ijk} \equiv \left(\frac{x_{ijkf}}{x_{ijks}}\right)^{\frac{1}{f-s}} - 1$$

For this sample of exports, we define the implied level of exports in year $s+1$ as $x_{ijk,s+1} \equiv \left(1 + g^C_{ijk}\right) x_{ijk,s}$.

We face a different problem if a country exports the product in year s, i.e., $\sum_j x_{ijks} > 0$, but no bilateral exports are reported between two regions in the industry in the start year, i.e., $x_{ijks} = 0$ for some $\{i, j, k, s\}$. To deal with this problem, we define the average growth rate in exports due to new export destinations as

$$g^N_{ik} \equiv \left(1 + \frac{\sum_{j \in \mathcal{N}_i} x_{ijkf}}{\sum_j x_{ijks}}\right)^{\frac{1}{f-s}} - 1, \tag{A.1}$$

where $\mathcal{N}_i$ is the set of new export destinations, which are defined to be the observations satisfying $x_{ijks} = 0$ and $x_{ijkf} > 0$. In this case, we set the annualized level of exports to new destinations in $s+1$ as $x_{ijk,s+1} \equiv \left(1 + g^N_{ik}\right)^{-(f-s-1)} x_{ijkf}$. In other words, we set the counterfactual amount of exports to new destinations in year $s+1$ equal to the observed amount of exports in year f (x_{ijkf}) deflated by the growth rate in exports due to extensive margin growth between years $s+1$ and f. With these annualized values for exports in hand, we can write the left-hand side of equation 6 as

$$\frac{\sum_j x_{ijkf} - \sum_j x_{ijks}}{\sum_j x_{ijks}} = \frac{\sum_j x_{ijk,s+1} - \sum_j x_{ijks}}{\sum_j x_{ijks}}, \tag{A.2}$$

and the left-hand side of equation 7 as

$$\frac{\sum_i x_{ijkf} - \sum_i x_{ijks}}{\sum_i x_{ijks}} = \frac{\sum_i x_{ijk,s+1} - \sum_i x_{ijks}}{\sum_i x_{ijks}}. \tag{A.3}$$

We then can apply the AW estimation procedure in equations 6 and 7 to estimate the γ_{ik}.

D Variables from External Sources

This section documents the variables we obtained from secondary sources and any changes we made to them. We discuss data from primary sources in the next sections.

D.1 Defining current high-income countries

We make a reference to "high-income" countries in the Introduction. We define a country as high income if its GDP per-capita (PPP adjusted) in 2022 is 50% or more of the US GDP per-capita, based on data from the World Bank (2024). Specifically, we use the variable "GDP per capita, PPP (current international dollars)."

D.2 Identifying the plurality language by country: Ethnologue (2023)

Reference Ethnologue, https://www.ethnologue.com/.

We identify the plurality language spoken by each country for the analysis examining the relationship between per capita-income and linguistic distance in Appendix Table A.1 and Appendix Figures A.1 - A.2. To do so, we obtain the modern (2023) plurality language spoken in each country from "Ethnologue".

D.3 Weeks to Learn a Language: Foreign Service Institute (2023)

Reference "Foreign Language Training - United States Department of State," U.S. Department of State, 03-May-2023. [Online]. Available: https://www.state.gov/foreign-language-traning/.

The Foreign Service Institute of the U.S. Department of State estimates the number of weeks required for an English native speaker to reach "General Professional Proficiency" in the language (a score of "Speaking-3/Reading-3" on the Interagency Language Roundtable Scale. We use this measure to proxy linguistic distance for the analysis examining the relationship between per capita-income and linguistic distance in Appendix Table A.1 and Appendix Figures A.1 - A.2.

D.4 Distance to U.K.: GeoDist Database (Mayer and Zignago, 2011)

We control for physical distance in the analysis examining the relationship between per capita-income and linguistic distance in Appendix Table A.1 and Appendix Figures A.1 - A.2. To do so, we use data from *Centre d'Etudes Prospectives et d'Informations Internationales* (CEEPI) which report different measures of bilateral trade distances (in kilometers) for 225 countries. Our measure of the distance between any two countries is the "dist" variable, which is calculated using the great circle formula. They compute internal distances by using the latitudes and longitudes of the most important cities/agglomerations (in terms of population). This means that the distance of a country to itself will never be zero; rather, the distance measure captures how far away major population centers within a country are from each other.

D.5 Historical income and population data: Maddison Project Database

The Maddison Project Database provides information on comparative economic growth and income levels over the very long run. We use the 2020 version of this database (Bolt and van Zanden, 2020), which covers 169 countries up until 2018. We use data on GDP per capita from this source for the analysis examining the relationship between per capita-income and linguistic distance in Appendix Table A.1 and Appendix Figures A.1 - A.2. Further, we also use this source to assign regions into income groups in the main analysis (Section 6).

Classifying regions as high-, medium- and low-income

We classify regions in our dataset by income level using the GDP per capita data from Maddison for 1870. To obtain this variable, we adopt the following steps:

1. The Maddison data uses modern country borders. We first map modern countries to the historic states they were part of in 1880-1914, which will match our trade data (e.g., Hungary and Austria map to Austria-Hungary).

2. The GDP per capita of a historical state that spans two or more modern countries is the simple mean of the GDP per capita of its constituent modern countries.

3. We rank regions by GDP per capita in descending order. Countries in the top third of this distribution are considered high income, countries in the middle third, middle income, and countries in the bottom third, low income.

Finally, we also use the Maddison data to estimate annualized population growth needed for constructing Figure 11.

Estimating annualized population growth

We use the 1870 and 1913 population data to estimate a region's population growth according to the following protocol:

1. Concord the modern countries in the Maddison database with the historic regions we use in this paper.

2. The population of a historic region for a given year is the sum of the population of the modern states that make it up.

3. Compute annualized population growth

$$\text{Annualized Population Growth}_i = \left(\frac{\text{Population}_{i,1913}}{\text{Population}_{i,1870}}\right)^{\frac{1}{1913-1870}} - 1$$

The Maddison Project does not report data for the Russian Empire during this time period; we complement the database by using the Russian population estimates for 1880 and 1910 from Mitchell (1975).

D.6 Historical Italian trade data: Federico et al. (2011)

We obtain Italian trade data for 1880, 1885, 1905, and 1910 from Federico et al. (2011). This dataset harmonizes historical trade records from Italian customs between 1862 and 1950 by concording the different product lines to SITC codes. The source reports bilateral trade at the product level between Italy and its ten biggest trading partners.

D.7 Historical Belgian trade data: Huberman et al. (2017)

We obtain the Belgian bilateral product-level trade data for 1880, 1885, 1905, and 1910 from Huberman et al. (2017). They use the *Tableau générale du commerce extérieur* published by the Belgian government as their primary source and concord product lines to SITC codes. The authors record trade *in manufacturing* at five-year intervals between 1870 and 1910. In 1900, 50% of Belgian exports and 20% of imports were in manufacturing.

D.8 Historical Japanese export data: Meissner and Tang (2018)

We obtained bilateral product level Japanese export data at five-year intervals between 1880 and 1910 from Meissner and Tang (2018). This dataset was constructed from the trade statistics volumes published by the Japanese Ministry of Finance. The authors concorded product lines to SITC codes.

D.9 French Energy Data: Chanut (2000)

We control for the intensity of steam usage of industries in our regressions. We construct this data based on French energy data that comes from Chanut (2000). We manually map French industries to SITC codes. We define the Steam Intensity of an industry as the ratio between the Steam Engine Horsepower of the industry over its Wage Bill. We define the wage bill as:

Wage Bill = (# of Male Workers)*(Avg. Male Hourly Wage) + (# of Female Workers)*(Avg. Female Hourly Wage) + (# of Child Workers)*(Avg. Child Hourly Wage).

D.10 Historical Exchange Rates: Fouquin and Hugot (2016)

Our bilateral-product level trade data converts the value of exports and imports (reported in local currency) into current yen. We use data on annual exchange rates from the *Historical Bilateral Trade and Gravity Dataset (TRADHIST)* from which we obtain the yearly exchange rates for the 1870-1915. Specifically, they provide us the value of one unit of the local currency in pounds.

We calculate the exchange rate from Yen to Belgian francs, Italian lira and US dollars as follows:

$$\frac{£_t/X_t}{£_t/¥_t} = \frac{¥_t}{X_t}$$

where t refers to year and X to the local currency. The value that we obtain is the value of one unit of the local currency in yen.

E Bilateral Trade Dataset

Our master bilateral, product-level trade dataset is constructed from four main sources:

- American exports and imports in 1880, 1885, 1905 and 1910
- Belgian manufacturing exports and imports in 1880, 1885, 1905 and 1910
- Italian exports to and imports from top trading partners in 1880, 1885, 1905 and 1910
- Japanese exports and imports in 1875, 1880, 1885, 1905 and 1910

As noted in the previous section, the Belgian and Italian trade data, as well as the Japanese export data for most years, has already been digitized and concorded to SITC by others.

We digitized and concorded to SITC the U.S. trade data, Japanese import data, and Japanese export data for 1875. The U.S. data are digitized from yearly volumes of *Foreign Commerce and Navigation, Immigration, and Tonnage of the United States* published by the Treasury Department's Bureau of Statistics (1900). The Japanese trade data was sourced from the yearly volumes of *Annual Return of the Foreign Trade of the Empire of Japan* published by the Department of Finance (1916). From these volumes, we only use the tables from the "Quantity and Value of Commodities Imported/Exported from Various Countries" sections. We use the Meissner and Tang (2018) product-SITC mapping wherever possible for Japan and the U.S. to ensure consistency.

Japan and the U.S. kept detailed records of their trade with foreign countries between 1880 and 1910. Each entry tells us the name of the product, its quantity, units, transaction value, and year, as well as the exporting and importing countries. The construction of these data involves digitizing the records and harmonizing products and country names. To construct the harmonized dataset across different reporting countries, we convert all data to a common currency, harmonize country names, and deal with double reporting issues. The protocols we adopted are described in detail in the subsections below.

E.1 Harmonization of Countries

Country names are not standardized across reporters (Belgium, Italy, Japan, and the U.S.) and years. In order to make comparisons across years and countries, we standardized country names as follows:

1. We made a list of all the country names that appear in all of the trade books from the four reporters.
2. We grouped names that refer to the same country: e.g., Vietnam and French Indo-China both refer to the same political entity at the time.
3. We kept the group if it is used by at least three reporters in the 1880/5 *and* 1905/1910 books for each reporter.

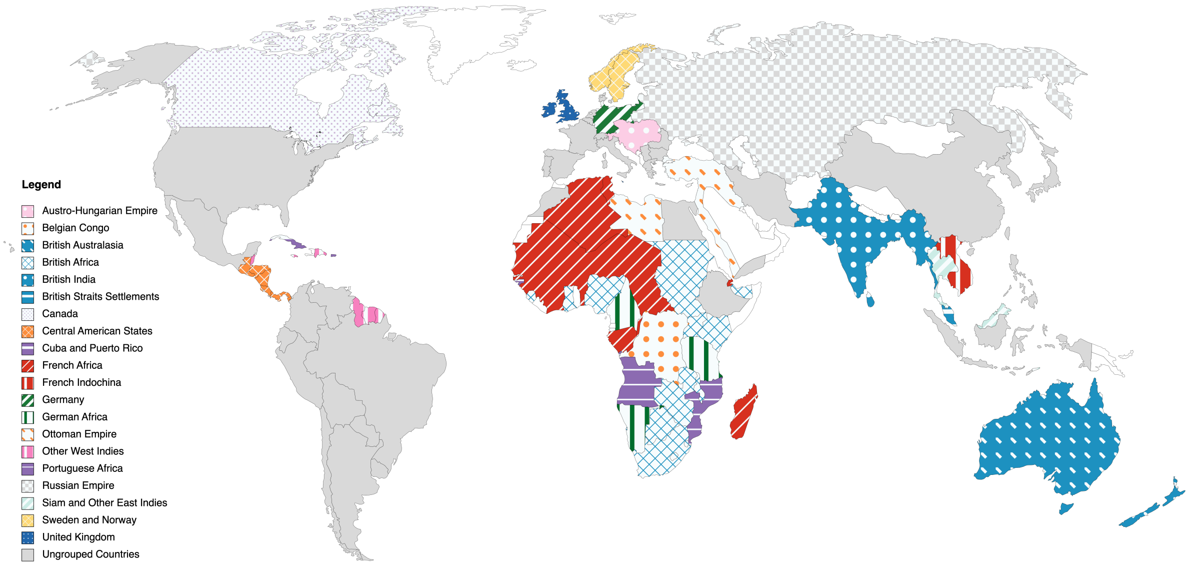
4. If the country group did not meet the previous requirement, then we try to build a regional group that does. For example, Honduras, Nicaragua, and Costa Rica do not have three reporters in the all the required years. If we group all Central American States together, this larger regional group meets our requirements.

5. If a country could not be grouped and did not meet the reporter-year requirement, then we dropped it.

6. If a region was too disaggregated, we dropped it. For example, Singapore and Hong Kong are their own separate categories, each with substantial volumes of trade in our dataset. If one country, in one year, reported “Hong Kong & Singapore," we dropped this observation.

The map illustrates how we grouped countries. Countries in grey were left as they were. We use the map of the world on the eve of World Word I (1914) as a baseline for our country groups.

Figure A.6: Country Groups



Created with historicalmapchart.net

Note: Colonies are grouped by imperial power and region (e.g. British Africa, French East Indies). All small, remote islands (e.g. Falklands) were dropped. Countries in white are missing from the dataset, countries in gray are reported as in the raw data. The remainder of the footnote reads from West to East on the map. The West Indies are grouped together with the exception of Cuba and Puerto Rico. British Honduras (although technically in Central America) is considered part of the West Indies due to its political affiliation with other British colonies in the Caribbean. The Ottoman Empire includes Libya, but not Algeria (which fell to the French in 1881). Taiwan is never directly mentioned in any trade statistics and not included in Japanese trade for the time period. Since each book either mentions French India or French Indochina, we conclude that French India refers to French Indochina, not to the French port cities in India. Thailand (then Siam) is grouped with other minor East Indies colonies such as Timor-Leste and British Borneo.

E.2 Double Reporting

Trade between reporting countries appears twice: once as exporters from the origin and secondly as imports by the destination. For all reporting countries except Belgium, we use their export data for their exports to reporting and non-reporting regions. Because Belgium does not report any trade data for non-manufacturing sectors, we use the reporting country's import data from Belgium to fill in these gaps. We use imports by reporting countries from non-reporting countries to construct the exports of non-reporting countries.

F Constructing the British Patent Relevance measure

F.1 Overview

We construct BPR by assuming that the similarity of text in books describing production techniques and the text of British patent data tells us the relevance of patents for that industry. In practice, we start with unigrams (i.e., single words such as "steam") and bigrams (i.e., two-word combinations such as "steam engine" or "steam engines"). We convert these into "terms" by stemming them and converting them into the terms "steam" and "steam engin". We also make use of two types of corpora. The first is the set of books describing production techniques in industry k, and the second is the are British patent synopses. Thus, we have a corpus for each industry and a separate corpus for the patents. To measure the relevance of the British patent corpus to industry k's corpus, we weight each term's frequency in a book by the total number of books divided by the number of books containing the term, i.e., we compute the Term Frequency-Inverse Document Frequency (TF-IDF). For each industry, we build a TF-IDF vector that characterizes its vocabulary (where each element is the TF-IDF of a term); we also build a TF-IDF vector for patents. Finally, we compute British patent relevance of industry k as the cosine similarity between the vector of TF-IDFs for industry k and that for the set of British patent synopses. We explain each of these steps in detail below.

F.2 Building the Terms

To build a term, we start with n-grams, we implement the following steps:

1. We split the raw text into sentences
2. We convert the words in the sentence to lower case, stem the words, replace UK spelling with US spelling
3. We turn each processed sentence in a sentence word list (where the position of a word on a list is the position it has in the sentence)
4. For each sentence word list, we split it into n-grams
5. We count the number of times an n-gram appears in the sentence and sum across sentences
6. Drop n-grams that include at least one stop word, (i.e., "a," "the," etc.)
7. Output a dataset with all the n-grams in the document and their count in the corpus

Example

1. **Text** "A stemmer for English operating on the stem cat should identify such strings as cats, catlike, and catty."
2. **Sentence** "A stemmer for English operating on the stem cat should identify such strings

as cats" "catlike" "and catty"

3. **Processed Word List** "a stemmer for english oper on the stem cat should identifi such string as cat" "catlik" "catti"

4. **Unigrams** "a" "stemmer" "for" "english" "oper" "on" "the" "stem" "cat" "should" "identifi" "such" "string" "as" "cat" "catlik" "catti"

5. **Unigrams without Stopwords** "stemmer" "english" "oper" "stem" "cat" "should" "identifi" "string" "cat" "catlik" "catti"

6. **Final Unigrams with Count** "stemmer" 1 "english" 1 "oper" 1 "stem" 1 "cat" 2 "should" 1 "identifi" 1 "string" 1 "catlik" 1 "catti" 1

F.3 Focusing on Jargon

Many unigrams and bigrams are not technical jargon. In order to focus our analysis on jargon, we drop unigrams and bigrams that are commonly used. We use the Bible to identify commonplace non-technical words that are necessary to write a coherent text but are not helpful in defining an industry's technical vocabulary. We use the 1885 King James Bible because it uses the common, non-technical nineteenth-century words and phrases. We define Biblical words as the 1,000 words occurring with the highest frequency in the Bible. However, if one of these words is used in the description of an SITC keyword, we do not count it as a Biblical word. For example, the stemmed word "brea" is a top 1,000 word in the Bible, but it also happens to be a keyword in the SITC description for cereal products.

F.4 Formally Defining TF-IDF

The term frequency (TF) measure is the count of instances a term appears in a corpus, divided by the number of terms in the corpus. The formula for the TF of term τ in corpus c is

$$\text{TF}(\tau, c) \equiv \frac{F_{\tau,c}}{\sum_{\tau' \in c} F_{\tau',c}} \tag{A.4}$$

where $F_{\tau,c}$ is the raw count of τ in c; and $\sum_{\tau' \in c} F_{\tau',c}$ is number of terms in the corpus. The inverse document frequency (IDF) is a measure of how common or rare a word is across all documents. The rarer the word, the higher the IDF score. We define the IDF for term τ in all corpora C (i.e., the complete collection of books) as

$$\text{IDF}(\tau, C) = \log\left(\frac{N}{N_\tau + 1}\right) \tag{A.5}$$

where N is the total number of books in C; N_τ is number of books in the corpus where the term τ appears.

The TF-IDF is then

$$\text{TF-IDF}(\tau, c, C) = \text{TF}(\tau, c) \cdot \text{IDF}(\tau, c) \tag{A.6}$$

We remove any n-grams that include words in the description of the SITC categories from the sample before estimating the cosine similarities. For example, removing the unigram "cotton" ensures that books describing how to grow cotton are not coded as part of the technology to spin cotton yarn.

Comparing the Vocabulary of Industries and Patents

We define the British Patent Relevance of industry k as the similarity between the TF-IDF vector representation of vocabulary for industry k and patent vocabulary. We use cosine similarity to measure the similarity between the two vectors. If an industry uses the same words at the same frequency as the patent book, then the vectors are the same, and we conclude that British patents are very relevant in the industry. If there is no overlap in words, then the similarity score is low, and we conclude that British Patents are not relevant. See the main text (equation eq:BPR) for the formal definition of cosine similarity.

F.5 Data Sources

All data (unless otherwise specified) was accessed through HathiTrust.

British Patents All patent text comes from the second edition of *Subject-Matter Index of Patent of Invention From March 2, 1617, to October 1, 1851 Parts I (A to M) and II (N to W)*, published by Woodcroft (1857).

Industry For each industry (as defined by SITC-3) we hand-curated a list of books and sections of nineteenth century books relevant in describing the production process of the goods in the industry.

Bible (1885) English Revised Version of the Bible.[1]

G New Japanese Words in the Meiji Period

We utilize the etymology of Japanese words based on the revised edition of *Nihon Kokugo Daijiten* [The Unabridged Dictionary of the Japanese Language], published by Shogakukan (2006). Importantly, it includes the title and year of publication of the Japanese document in which each word is believed to have been first used. We obtained the digitized data for this dictionary from Kotobank.[2] The number of new words by year can be seen on Figure 7.

H Technical Books in the Top World Languages (1800-1910)

H.1 Overview

We report the source libraries for our data on technical books in Table A.10. We tried, where possible, to scrape national libraries. If we could not find a scrapable national library for a language (such as Arabic and Russian), we scraped WorldCat, an online catalog of thousands of libraries worldwide covering dozens of languages. Scraping national library catalogs has an advantage over using WorldCat as the latter source sometimes overstates the number of books because different libraries sometimes report book titles differently (e.g., slight variations in titles or author names).

We minimized the number of possible duplicates by removing spacing and punctuation in book titles and dropping any duplicated book titles published in the same year. In order to minimize the role played by reprints of the same book, we also dropped any duplicates arising from books (possibly published in different years) with the same book ID. Importantly, the number of books

[1] Wikepedia article Revised Version of the Bible: https://en.wikipedia.org/wiki/Revised_Version

[2] Kotobank: https://kotobank.jp/dictionary/nikkokuseisen/

reported for four of our five top codifying languages, French, English, German, and Japanese (but not Italian), were from national libraries, so we can be confident that there is minimal double counting in these book totals.

If we could scrape a national library or WorldCat, we made a judgment call about which source was better. If we saw that for a non-top-4 codifying language, there were more *genuine* technical books than we could find in a national library, we opted for the number from WorldCat. For example, the national libraries of Portugal and Spain have very few technical books in their catalogs relative to the libraries in WorldCat, so we opted to use WorldCat for these languages. Because of the duplication issue in WorldCat and the fact that WorldCat allows us to scrape many libraries for each language, our use of national libraries for English, French, German, and Japanese likely causes us to understate the concentration of technical books in these languages.

We scraped the number of technical books for 33 languages, which include all of the 20 most spoken native languages on earth.[3] We define the set of books comprising technical knowledge as those with a subject classified as applied sciences, industry, technology, commerce, and agriculture. For our purposes, we exclude books on theoretical technical knowledge, such as books in the hard sciences or in medicine.

[3] We assume that if someone speaks Yue or Wu Chinese, they can read Mandarin Chinese, given that these languages all use the same characters.

Table A.10: Catalogs Scraped

Library	Catalog	Languages	Years	Classification System	Tech Topics
Bibliothèque Nationale de France	Link	French	1500-1930	Universal Decimal Classification	Applied Sciences and Technology (6)
Deutsche Nationalbibliothek	Link	German	1500-1930	Dewey Decimal Classification	Technology (600)
National Diet Library	Link	Japanese	1500-1930	Nippon Decimal Classification	Technology (500) Industry (600)
Korean National Library	Link	Korean	0022-1980	Dewey Decimal Classes	Technology and Engineering (600)
Library of Congress	Link	English	1500-1930	Keyword Search	Hand-constructed
National Library of India	Link	Bengali Hindi Marathi Tamil Urdu	1500-1980	Only has three options	Non-Fiction Manually picked tech books.
Shanghai Library	Link not accessible	Chinese	1500-1980	Chinese Library Classification System	Agriculture (S) Industry (T) Transportation (U)
WorldCat	Link	Arabic Bulgarian Croatian Czech Danish Dutch Greek Hebrew Indonesian Italian Norwegian Persian Polish Portuguese Romanian Russian Spanish Swedish Thai Turkish Ukranian Vietnamese	1800-1930	Subject filter in advanced search	Hand-constructed

H.2 Search Filters

1. **Format:** We only search for books. No images, periodicals, articles, or newspapers.

2. **Language:** We always specify the language of the text. For example, when searching the National Diet Library, we only look for books written in Japanese.

3. **Publication Year:** 1500-1930

4. **Subject:** We always search by subject.

   - We search by subject code, if possible. Otherwise, we manually picked technical books.
   - If subject codes are not available, we use subject keywords. To do this, we first find the underlying subject classification system used by the library (e.g., Dewey Decimal Classification) to get the descriptions of the subject codes we want.