Online Appendix

A Government Expenditures

Figure A.1 shows that the imposition of the land tax enabled the early Meiji government to finance enormous investments in codification and technical absorption.

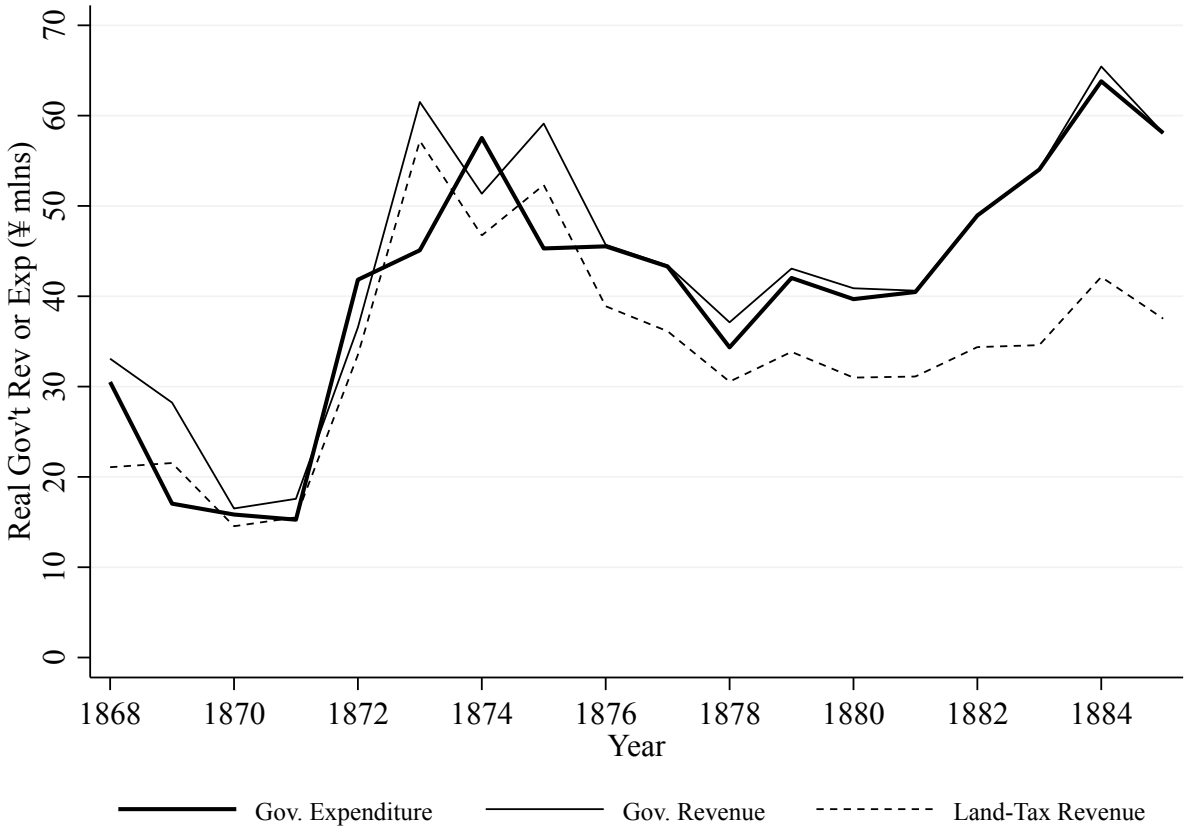


Figure A.1: Japanese Government Expenditure

Note: Government expenditure and revenue data are from Toyo Keizai Shimposha (1926) *Meiji Taisho Zaisei Shoran [Meiji and Taisho Financial Details]*, Toyo Keizai Shimposha: Tokyo, pp. 2 and 640. Before adopting the Gregorian calendar in 1873, Japanese fiscal years varied in duration and did not align perfectly with Western ones, but the mapping to Western years is approximately correct. These are deflated by the Wholesale Price Index from Ohsato, Katsuma (ed., 1966) *Hundred-Year Statistics of the Japanese Economy*, Statistics Dept., The Bank of Japan: Tokyo, p. 76.

As a result of Japan's impressive ability to raise government revenues, by 1884, Japanese government revenues equaled 83.1 million yen. By contrast, the Chinese government in 1884, still recovering from the chaos of the Opium Wars and Taiping Rebellion, could only raise 114 million yen even though China had ten times Japan's population.[1]

[1]Wong (2012) reports that Chinese tax revenue in 1884 was 77 million silver taels. We performed the

B Mastery of IR Technologies Required for Developing Newer Technologies: Historical Evidence From Japan

In Section 4.1, we argued that Japan needed to absorb IR technologies before it could master newer technologies available at the technology frontier by 1880. Here, we present additional historical evidence for Japan, which suggests that industrial development around 1880 was not nearly sufficiently mature for machine building and the other related sectors to emerge (Suzuki, 1999; Masanori, 2022). For machine-building in particular, interchangeable parts were a complex technical feat requiring a high level of precision and quality from related sectors (e.g., castings, steel). Until 1910 (when our sample period ends), Japanese industry did not possess these capabilities. In fact, consistent with the literature on the big-push (e.g., Murphy et al. (1989)) and sectoral linkages (Hirschman, 1958), it was necessary for Japan to master IR technologies before it could become competitive in sectors such as machine-building that required high-quality inputs (such as bolts, fittings, and standardized parts) and the knowledge acquired from mastering the first set of technologies.

For example, technicians from the cotton spinning industry assisted in the development of Toyoda's power loom (a mechanized machine for weaving) in 1909 (Suzuki, 1999). The knowledge acquired in mastering cotton spinning allowed the Japanese industry to move into machine building. Finally, we note that this discussion provides a micro-foundation for the technology adoption lags literature (Comin and Hobijn, 2010). Japan adopted interchangeable parts with a substantial lag relative to the West (where interchangeable parts were an integral part of the American System of Manufacturing that emerged in the early to mid-19th century), because the Japanese domestic economy was missing complementary capabilities until after the turn of the 20th century.

C Productivity Growth

C.1 Estimating Productivity Growth

In this section, we demonstrate how to utilize trade data to construct a global database that enables us to estimate productivity growth at the region-industry level. Here, we explain how we estimate productivity growth for our set of regions. The basic intuition for this procedure is based on the Ricardian model of trade. In the canonical two-country, two-good version of this model, knowing the relative labor productivities of the cloth and wine industries in England and Portugal tells us which country will export which product. The simple Ricardian model cannot be applied to data because the prediction that a country cannot import a good it exports is patently false. Costinot et al. (2012) solve this problem using the theoretical setup of the Eaton and Kortum (2002) model. In their model, each industry (k) in exporter (i) is composed of a continuum of varieties (goods) each produced based on a random productivity draw (z), whose mean rises with the "fundamental" productivity in the industry, z'_{ik}, where average industry productivity is a linear function of z'_{ik}. Thus, if a country has a high average productivity in some industry, it will tend to be the low-cost supplier of more varieties in that industry and therefore export more. Since there is a monotonic relationship between productivity and the value of exports, we can invert this relationship to obtain an estimate of productivity by observing the level of exports. Costinot et al. (2012) show that the relationship between exports from i to j in any period t (x_{ijkt}) and fundamental productivity in an

currency conversion in two ways. The number in the text uses the exchange rate series from (Fouquin and Hugot, 2016) of 1.39. We obtain a similar estimate if we convert silver taels into yen by noting that an 1867 Shanghai silver tael contained 36.0 grams of silver and an 1876 silver yen coin contained 24.3 grams of silver, according to https://en.numista.com. This implies an exchange rate of 1.48 yen per tael.

industry at time t (z'_{ikt}) is linear in logs and can be written as

$$\ln x_{ijkt} = \gamma'_{ijt} + \gamma'_{jkt} + \theta \ln z'_{ikt} + \epsilon'_{ijkt}, \tag{A.1}$$

where γ'_{ijt} is an importer-exporter fixed effect; γ'_{jkt} is an importer-industry fixed effect; $\theta > 0$ is the Fréchet scale parameter; and ϵ'_{ijkt} is an error term that captures how trade costs deviate at the industry-exporter-importer level from the exporter-importer average. The intuition for this formula is that the amount trade between two countries will depend on bilateral factors captured by γ'_{ijt} (such as bilateral distance, the relative sizes of the exporter and importer, etc.), industry demand conditions in the importer captured by γ'_{jkt}, and relative productivity of the exporter in the sector (x_{ijkt}). We could estimate $\theta \ln z'_{ikt}$ by regressing log bilateral exports on an ijt, jkt, and ikt fixed effects, but given the large number of zero trade flows, this would be biased.

Our path into solving this problem is to first note that our objective is to estimate not the level of productivity, but the change: $\gamma_{ikt} \equiv \theta \Delta \ln z'_{ikt}$. We estimate it by noting that we can first-difference equation (A.1) and rewrite it in terms of fixed effects:

$$\Delta \ln x_{ijk} = \gamma_{ij} + \gamma_{jk} + \gamma_{ik} + \epsilon_{ijk}, \tag{A.2}$$

where we have suppressed the time subscript and $\gamma_{\ell,m} \equiv \Delta \gamma'_{\ell,m}$ for any index (ℓ, m). Estimating this equation enables us to identify γ_{ik} and therefore $\theta \Delta \ln z_{ik}$ up to the choice of a normalization that pins down the reference exporter productivity, importer demand, and industry productivity.[2] This equation can be rewritten to yield

$$\Delta \ln x_{ijk} = \gamma_{jk} + \gamma_{ik} + \tilde{\epsilon}_{ijk}, \tag{A.4}$$

where variables without primes correspond to the first differences of variables with primes and $\tilde{\epsilon}_{ijk} \equiv \gamma_{ij} + \epsilon_{ijk}$.

Estimation of equation (A.4) requires us to drop observations whenever the initial bilateral export flow in a exporter-importer-industry tuple is zero, which is problematic because a large amount of nineteenth-century export growth was due to exporters expanding their set of export destinations over time. This can bias estimates of productivity growth based on a log-difference specification downwards because it cannot account for growth due to the extensive margin. Amiti and Weinstein (2018) [AW] propose an alternative estimation approach that corrects this problem.

Their estimator is closely related to weighted least squares. In particular, if there are no zeros in the export data, the AW estimates will match those obtained using weighted least squares with lagged export weights. A unique property of the AW estimates of γ_{jk} and γ_{ik} is that they aggregate to match the growth rate of total exports in every region-industry in which the industry's *aggregate* growth rate is well defined: i.e., the region initially has positive exports to at least one country in the industry. Similarly, the estimates aggregate to match region-industry import levels as long as a region has positive imports from at least one country in the industry in the initial period. Thus, an export-weighted average of the γ_{jk} and γ_{ik} will match *total* export growth in each country

[2] One can see this by noting that equation (A.2) can be rewritten as

$$\Delta \ln x_{ijk} = (\gamma_{ij} + \gamma_i + \gamma_j) + (\gamma_{jk} + \gamma_k - \gamma_j) + (\gamma_{ik} - \gamma_i - \gamma_k) + \epsilon_{ijk}, \tag{A.3}$$

where γ_i, γ_j, and γ_k are arbitrary normalization constants that define the baseline exporter productivity, importer demand, and industry productivity.

and industry.[3] One can formally see that the AW estimator will have this property by writing down the moment conditions used to obtain the estimates. In particular, the estimates will satisfy two types of moment conditions. First, the estimates aggregate to match total exports in every exporter-industry observation i:

$$\frac{\sum_j x_{ijk,t} - \sum_j x_{ijk,t-1}}{\sum_j x_{ijk,t-1}} = \gamma_{ik} + \sum_j \frac{x_{ijk,t-1}}{\sum_\ell x_{i\ell k,t-1}} \gamma_{jk}, \tag{A.5}$$

where we have added a time subscript, t, to be clear about how time differences are constructed from changes in levels. The left-hand side of the moment condition equals the growth rate of *total exports* in sector k from exporter i, and the right-hand side is the sum of the exporter fixed effect (γ_{ik}) and a bilateral export weighted average of the importer fixed effects (γ_{jk}). This condition, therefore, ensures that an export-weighted average of the parameters aggregates to match total exports. Second, the estimates will aggregate to match total imports in every importer-industry observation j because they impose a second moment condition:

$$\frac{\sum_i x_{ijk,t} - \sum_i x_{ijk,t-1}}{\sum_i x_{ijk,t-1}} = \gamma_{jk} + \sum_i \frac{x_{ijk,t-1}}{\sum_\ell x_{\ell jk,t-1}} \gamma_{ik}. \tag{A.6}$$

Here, the left-hand side of this moment condition is the growth rate of *total imports* in sector k by importer j, and the right-hand side is the sum of the importer fixed effect (γ_{jk}) and a bilateral export weighted average of the exporter fixed effects (γ_{ik}). Since the estimates satisfy these two moment conditions, the AW estimates aggregate to match the growth of exports and imports in every region for each industry.

Once we obtain the estimates of γ_{ik} and γ_{jk}, we run the following regressions to impose normalizations that lead to a meaningful decomposition of global trade patterns:

$$\gamma_{ik} = \gamma_i + \gamma_{1k} + \tilde{\gamma}_{ik}, \tag{A.7}$$

and

$$\gamma_{jk} = \gamma_j + \gamma_{2k} + \tilde{\gamma}_{jk}, \tag{A.8}$$

where $\tilde{\gamma}_{ik}$ and $\tilde{\gamma}_{jk}$ are regression residuals. This normalization choice has several useful properties. First, γ_i tells us the growth in exports resulting from shifts in exporter characteristics (e.g., productivity or size). Second, $\tilde{\gamma}_{ik}$, the "comparative-advantage" component of export growth, corresponds to the growth in exports due to shifts in productivity that are orthogonal to changes in exporter factors (i.e., γ_i) and changes in industry factors (γ_{1k}).[4] Since the former captures shifts in productivity at the national level and the latter captures the impact of comparative advantage

[3] We also considered using the Poisson pseudo-maximum likelihood (PPML) estimator. However, one well-known issue with PPML is that it often fails to converge in datasets with many zeros like ours (Santos Silva and Tenreyro, 2010). While the AW estimator only required us to drop country-industry observations where there were no exports to or imports from *any country* in the initial period, the PPML estimator did not converge unless we used data for countries with at least two export destinations or two import sources in each industry. As a result, while the AW procedure produced 1,358 productivity estimates based on 6,216 observations, the PPML estimator only converged on a subsample that was 36.5 percent as large. The PPML estimator only produced 38 percent as many productivity growth estimates as the AW estimator.

[4] Although we do not use the other normalization constants, we can recover them. $\hat{\gamma}_k \equiv \hat{\gamma}_{1k} + \hat{\gamma}_{2k}$ is the shift in exports that can be attributed to movements in industry k's characteristics (e.g., global productivity growth in k or global demand for k). Similarly, γ_{ij} can be recovered by regressing $(x_{ijk,t}/x_{ijk,t-1} - 1 - \hat{\gamma}_{ik} - \hat{\gamma}_{jk})$ on ij fixed effects.

on export growth, $\tilde{\gamma}_{ik}$. In the Costinot et al. (2012) model, $\gamma_{ik} \equiv \theta \Delta \ln z'_{ikt}$, which enables us to define define $\Gamma_{ik} \equiv \tilde{\gamma}_{ik}/\theta$ as the change in exporters i's comparative advantage in industry k (i.e., the shift in productivity that cannot be explained by relative growth in industry k's productivity in all countries or relative productivity growth in the exporting country.[5]

In the following sections, we estimate γ_i and Γ_{ik} to understand patterns of productivity growth worldwide. We implement this methodology on annualized trade growth rates for the sample period (1880-1910), so our estimates correspond to averaged annual productivity growth rates. We show how to construct annualized rates in appendix C.2. All results reported below refer to annualized estimates.

C.2 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

$$g^N_{ik} \equiv \left(1 + \frac{\sum_{j \in \mathcal{N}_i} x_{ijkf}}{\sum_j x_{ijks}} \right)^{\frac{1}{f-s}} - 1, \tag{A.9}$$

[5] We follow Eaton and Kortum (2002) and set $\theta \equiv 8.28$. The choice of θ does not qualitatively affect any of our results; it just raises or lowers all countries' productivity growth proportionally.

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 $x_{ijk,s+1} \equiv \left(1 + g_{ik}^N\right)^{-(f-s-1)} x_{ijkf}$. 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 A.5 as

$$\frac{\sum_j x_{ijkf} - \sum_j x_{ijks}}{\sum_j x_{ijks}} = \frac{\sum_j x_{ijk,s+1} - \sum_j x_{ijks}}{\sum_j x_{ijks}}, \tag{A.10}$$

and the left-hand side of equation A.6 as

$$\frac{\sum_i x_{ijkf} - \sum_i x_{ijks}}{\sum_i x_{ijks}} = \frac{\sum_i x_{ijk,s+1} - \sum_i x_{ijks}}{\sum_i x_{ijks}}. \tag{A.11}$$

We then can apply the AW estimation procedure in equations A.5 and A.6 to estimate the γ_{ik}.

C.3 Productivity Growth Results

Section 3 examined Japan and other regions' economic performance using the raw trade data. Here, we utilize the methodology developed in the Section C.1 to provide the first systematic estimates of productivity growth for many regions in the late nineteenth and early twentieth centuries. Our normalization choice implies that productivity or anything that shifts exporter i's exports conditional on demand conditions will be captured by our estimate of γ_i. We can interpret $\hat{\gamma}_i - \hat{L}$, where $\hat{L}$ is the annual population growth rate, as a measure of exporter productivity, i.e., how much exports in country i grew after controlling for demand conditions and population growth. Figure A.2 plots the annualized per capita shift in export supply net of population growth relative to the value for the US, i.e., $\hat{\gamma}_i - \hat{L}_i - (\hat{\gamma}_{US} - \hat{L}_{US})$.[6] shows that the patterns are similar if we do not account for differences in population growth.

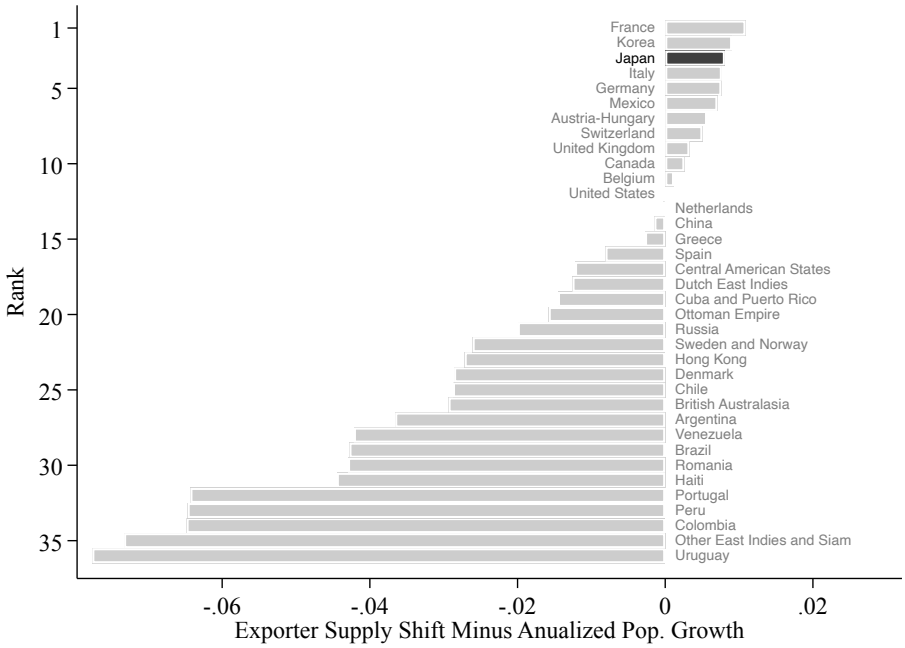
Reassuringly, the ranking of economies broadly aligns with what economic history teaches us about this period. France, Korea, Japan, Germany, Mexico, Italy, Austria-Hungary, Switzerland, the United Kingdom, Canada, Belgium, and the US show robust growth in their export supply shifter. In contrast, economies such as those of Portugal, Peru, Colombia, and Uruguay exhibit weak performance. Notably, Japan's export-supply shifter ranks third, confirming that its economy experienced some of the highest export productivity growth globally during this period. Notice that our estimates also suggest that Korea had high productivity growth (alongside Japan), which may be related to the fact that Japan forcibly opened Korea in 1876, and although nominally independent, the Japanese "reform[ed]" the Korean government and military administration by introducing to the country the kinds of measures that Meiji Japan itself had undertaken" (Iriye, 2007, p. 769)). Our result is consistent with the idea that the Meiji reforms may have also raised productivity in Korea.

Next, we examine the extent to which productivity growth was biased towards manufacturing. We regress the comparative-advantage component of productivity growth, Γ_{ik}, on broad industry dummies:

$$\Gamma_{ik} = \beta_i^{\text{Agg}} \times I_k^{\text{Agg}} + \beta_i^{\text{Mfg}} \times I_k^{\text{Mfg}} + \beta_i^{\text{Min}} \times I_k^{\text{Min}} + \epsilon_{ik} \tag{A.12}$$

[6] Appendix Figure A.3

Figure A.2: Relative Annualized Per Capita Exporter Supply Shifter by Exporter

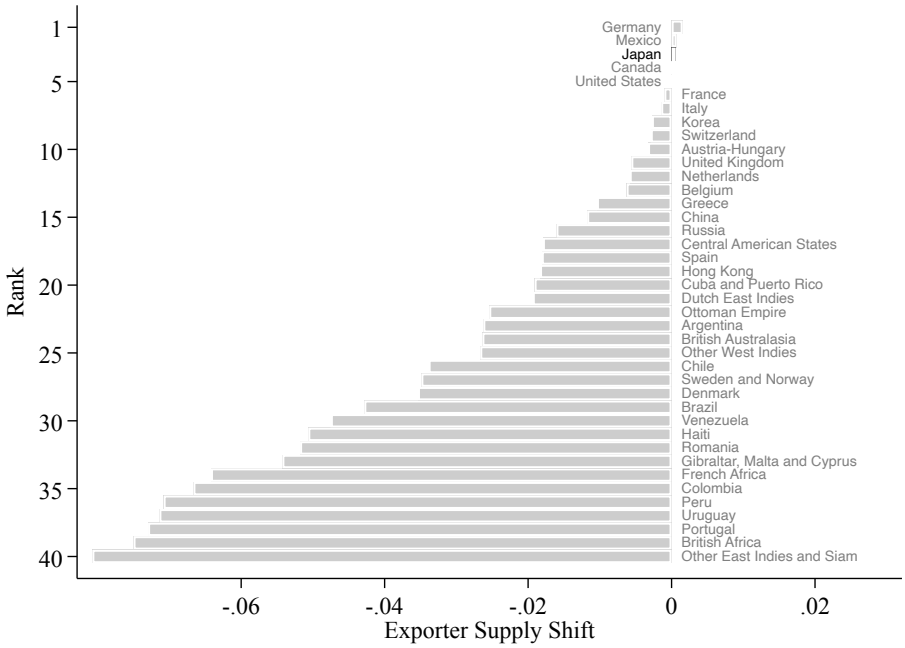


Note: Annualized per-capita exporter supply shifts are defined relative to the US, i.e., they are defined as $\hat{\gamma}_i - \hat{L}_i - (\hat{\gamma}_{US} - \hat{L}_{US})$. Annual population growth is computed between {1870,1880} and 1913 using the Maddison data (see Appendix H.5 for details).

where I_k^{Agg}, I_k^{Mfg}, and I_k^{Min} are dummies that are one if sector k is in agriculture, manufacturing, or mining, respectively; and β_i^{Agg}, β_i^{Mfg}, and β_i^{Min} are parameters that measure the average growth rate of comparative advantage for exporter i in agriculture, manufacturing, and mining. In words, $(\beta_i^{\text{Mfg}} - \beta_{US}^{\text{Mfg}})$ tells us how fast productivity in manufacturing grew in exporter i relative to the US after controlling for its average growth and the average growth in world manufacturing. Figure A.4 reports the results from this exercise for countries in which the manufacturing share of exports in 1880 was not trivial. While Portugal and Hong Kong exhibit strong shifts in comparative advantage towards manufacturing, the results in Figure A.2 indicate that these economies had low overall rates of productivity growth, implying that their relatively strong performance in manufacturing was offset by their low overall productivity growth. The next seven countries (Japan, Belgium, Mexico, Italy, the UK, the US, and Canada) are all examples of regions that industrialized during this period, exhibiting rapid productivity growth and exceptionally high relative productivity growth in manufacturing.

Our structural estimates of industry productivity growth in this period confirm that Meiji Japan's economic performance was exceptional. Average productivity growth was high in international comparison and shifted strongly towards manufacturing. This result supports the idea that Japan's unparalleled shift towards specialization in manufacturing (Figure 2) was driven by productivity growth biased towards manufacturing—that is, shifting Ricardian comparative advantage.

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.

D External Validity: Codification and Economic Performance in Other Periphery Economies Before WW1

In Section 7, we examined the influence of the Meiji model for Korea and China in the postwar era. Here, we are interested in further exploring the assertion that codification in the vernacular was a necessary, but not sufficient, condition for development in the late 19th century, consistent with other periphery economies' experience at this time.

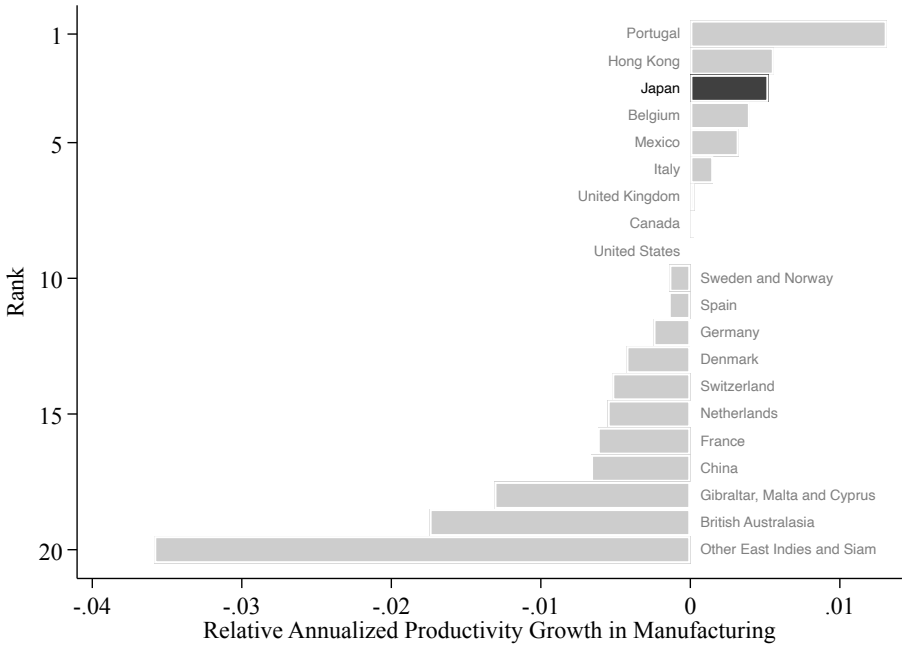
In Section 5, we showed econometric evidence that suggested that other periphery economies did not experience similar patterns of development. Here, we complement this evidence with historical evidence contrasting the experience of Japan with that of British India and Late Imperial Russia. Both have been the subject of influential case studies in industrial development (e.g., Gerschenkron (1962); Clark (1987), and each built up a sizeable modern, factory-based manufacturing sector by the eve of World War 1.[7]

We begin with British India, where—despite an early start compared to other periphery economies—Indian industry was quickly outcompeted by Japan in key sectors. By the 1930s, India had become the largest market for Japanese cotton cloth, even under a protective import tariff (Mass and Lazonick, 2013). Evidence shows that labor productivity and total factor productivity in Indian factories were especially low (Gupta and Roy, 2017), suggesting persistent difficulties in operating new technologies efficiently.

Turning to data on codification, Figure 5 shows that in 1910, there were virtually no technical

[7]For example, Russia and British India had the largest installed capacity in mechanized cotton spinning among periphery countries (U.S. House of Representatives, Tariff Board, 1912).

Figure A.4: Relative Annualized Productivity Growth in Manufacturing



Note: The plot presents our estimates of productivity growth in manufacturing relative to the US, i.e., $(\beta_i^{\text{Mfg}} - \beta_{US}^{\text{Mfg}})$. β_i^{Mfg} is estimated in equation A.12 for regions in which the manufacturing sector's export share in 1880 is at least 0.5% and for regions in which we can estimate productivity growth in at least five non-primary and five primary sectors.

books written in the major spoken languages: Hindi, Tamil, or Urdu. At first glance, British India therefore seems to be consistent with our assertion that codification in the vernacular was a necessary condition for development. While literate Japanese could read technical books in their language, Indians literate in local languages had no access to such material.

However, as a British colony, English was the lingua franca for higher education and technical instruction following the 1835 English Education Act. The real question, then, is whether Indians could access knowledge in English. Indian census data offers significant insight into this issue. The 1891 Census of India states that 537,811 people could read English, which was only 0.19 percent of the population (Government of India, 1893, p.224). Twenty years later, the 1911 Census reports that just 0.54 percent of Indians were literate in English (Government of India, 1913, p. 299). However, these figures likely overstate the actual number of Indians who could read English because they include foreign English speakers (e.g., British expatriates) who mostly lacked the ability to explain technical material in Hindi, Tamil, or Urdu. For example, the 1891 Census notes that only 386,032 Indians, or 0.14 percent of the population, could read English (Government of India, 1893, p.224), roughly the same as the percentage of Americans today who can speak Japanese: 0.15 percent (U.S. Census Bureau, 2022, p.3).

One can put the number of bilingual Indians into perspective by comparing it to the number of Japanese who could read Western technical manuals translated into Japanese. Given that Japan's population in 1891 was 41 million, and assuming a literacy rate of 40 percent, we estimate that approximately 16 million Japanese could read technical manuals. These numbers suggest that there were more than 40 Japanese people who could read technical manuals in Japanese for every

Indian person who could read them in English.

Of course, having forty times more people able to read engineering in Japan than in India might not have mattered if English speakers could easily share their knowledge across the language barrier in India. Although we lack concrete data on how difficult this was, Western and Indian historical accounts suggest that it was challenging for English speakers to communicate with those who could not understand them. Indeed, they argue that differences in training and literacy significantly contributed to the productivity gap between the Japanese and Indian textile industries. For example, Pearse (1930) study of the development of the Asian cotton textile industries notes that

> "Each [Japanese] firm has at least one engineer with university education and special textile engineering training. Some of the mill managers have passed through similar educational institutions, but all have at least graduated from one of the technical schools One notices everywhere the result of a good general education; the inside managers and foremen have had a sound training in technical schools, they have not grown up empirically in the mill; every mill girl reads and writes, and possesses general education quite on par with that of European countries. The foreman and general supervisors are specially trained in classes run by the combines. We are not dealing with labor as it exists in India, China, or South America" quoted in (Otsuka et al., 1988, pp. 84-85).

Similarly, Mehta (1954)'s 100-year history of the cotton textile industry in India emphasizes the communication and staffing challenges that arose from trying to use technology whose descriptions were written in English.

> "The difficulties of language [faced by English engineers] were unusually great, not only in relation to the workers but frequently also in relation to the employers and other members of the latter's office. The growth of other professions, namely, law, medicine and government service, generally precluded from the industry the extremely small number of Indians who had access to schools where English was taught. An exceedingly small number of Indians received their training in English technical institutes and factories. The capacity of the managing agents to ensure a high level of production on the basis of an informed judgment was extremely low in the first fifty years (i.e., from 1854 onwards). For one, the top technicians were Englishmen on whom direct control was extremely limited. Secondly, the managing agent was himself a novice in many cases in the art of management, not only of machines but also of men, and he was hardly fitted to achieve a proper control of production functions. (Mehta, 1954, pp. 101-108)

In light of the evidence above, the failure of British India to develop an internationally competitive industry aligns well with our narrative. With neither access to codified knowledge in spoken languages such as Hindi, Urdu, or Tamil, nor widespread literacy in English, Indians did not have access to codified technical knowledge.

Imperial Russia is another context which has been the subject of influential studies on late industrialization. Figure 5 shows the limited availability of technical books in Russian in 1910. Our theory would thus predict that Russia would struggle to develop an internationally competitive, modern industrial economy. Unfortunately, given the present state of knowledge, there is substantial debate in the literature about exactly how successful Russia's industrialization was during

this period (see e.g., Zhuravskaya et al. (2024) for a recent overview). This makes it essentially impossible to draw definitive conclusions about whether Imperial Russia industrialized without access to a level of codified knowledge comparable to Japan's.

However, if one examines the comparative performance of a flagship industry such as mechanized cotton spinning, there are important differences between the two countries. In particular, while both Japan and Russia had a sizeable domestic cotton textiles industry, a key difference was that Russia's industry developed behind a high tariff wall and predominantly served the domestic market (Gregg, 2020).[8] In fact, Gregg (2020, p. 162) characterizes the Russian cotton industry as having achieved "a worldwide intermediate case of industrial development." Consistent with the narrative of modest progress in the industry, Clark (1987) argues that Russian textile workers who migrated to New England were 54 percent as productive as English textile workers. Contrast this with Japan, where cotton textiles became an important exported commodity during our sample period. That is, while Japanese cotton textile producers were sufficiently productive to compete in international export markets, there is no evidence that this was the case for Russian producers on the eve of World War I.

In summary, the qualitative evidence for British India paints a consistent picture with the econometric evidence presented in the main text. The lack of codified knowledge in major spoken languages, combined with a low proportion of English speakers among the native population, kept knowledge access costs high in British India relative to Japan. This may be an important reason why, despite a generous head start, Japan rapidly outperformed British India in modern industries. We know much less about both codification and development in Imperial Russia. While future scholarship on Russian industrial performance before World War 1 may lead us to revise our conclusions, given the current state of knowledge, the evidence suggests that Russia is also consistent with a country in which high technology access costs precluded the emergence of an internationally competitive industrial sector.

[8]Japan was prohibited from enacting protective tariffs during this period due to the unequal treaties it was forced to sign.

E Additional Tables

Table A.1: Linguistic Distance from English and GDP

	Log GDP per Capita					
	(1)	(2)	(3)	(4)	(5)	(6)
	1870	1913	2018	1870	1913	2018
Log Physical Distance between Country and the UK	-0.170***	-0.207***	-0.237***	-0.248***	-0.315***	-0.323***
	(0.058)	(0.064)	(0.066)	(0.054)	(0.065)	(0.072)
Number of Weeks Required to Learn the Plurality Language	-0.010***	-0.013***	-0.008*	-0.005**	-0.007***	-0.003
	(0.002)	(0.003)	(0.004)	(0.002)	(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-speaking native will take to attain "Professional Working Proficiency" in the country's plurality language is estimated by the U.S. Department of State's Foreign Service Institute. See Appendix H for data construction and sources. Robust standard errors are in parentheses. $^{*}p < 0.10$,** $p < 0.05$,*** $p < 0.01$.

Table A.2: Manuals with the Most Copies Held by the Imperial College of Engineering Library

Category	**Author**	**Title**	**Copies**
Mathematics	Wilson	*Elementary Geometry*	340
	Todhunter	*Trigonometry for Beginners*	234
	Wilson	*Algebra for Beginners*	192
Civil Engineering	Unwin	*Elements of Machine Design*	71
	Rankine	*Applied Mechanics*	55
	Rankine	*Manual of Civil Engineering*	55
	Perry	*Treatise on Steam*	48
	Goodeve	*Elements of Mechanism*	34
Mining and Mineralogy	Egleston	*Hydraulic Mining in California*	62
	Milne	*Notes on the Ventilation of Mines*	47
	Lyman	*Reports of Progress for the First Year of the Oil Surveys*	30

Source: Reproduced from Meade (2022), Table 1, p. 12.

Table A.3: Summary Statistics

Variable	N	Mean	SD	p25	p50	p75
Change in exporter's i comparative advantage in industry k (Γ_{ik})	1246	0.00	0.04	-0.01	0.00	0.02
Change in Japan's comparative advantage in industry k ($\Gamma_{\text{Japan},k}$)	56	0.00	0.05	-0.01	0.01	0.02
Exporter's Industry Growth Rate	1397	-0.10	0.38	-0.05	0.03	0.09
Exporter's Industry Growth Rate in Japan	71	-0.05	0.37	-0.02	0.04	0.15
Britsh Patent Relevance	125	0.05	0.09	0.02	0.04	0.06

Note: The estimation of Γ_{ik} is detailed in Appendix Section C. Exporter's Industry Growth Rate is the annualized export growth rate for each industry between {1880, 1885} and {1905, 1910}. The details on the construction of British Patent Relevance are in Appendix Section I.

Table A.4: 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	-0.104**	-0.027	0.011	0.028**	0.022***	0.020***	0.014**
	(0.049)	(0.020)	(0.017)	(0.011)	(0.008)	(0.007)	(0.006)
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: $^{*}p < 0.10$,$^{**}p < 0.05$,$^{***}p < 0.01$.

Table A.5: Annualized Export Growth and British Patent Relevance - British Colonies and Steam Intensity

	Export Growth	
	(1)	(2)
BPR × Japan	0.121***	0.123**
	(0.033)	(0.052)
BPR × Not Japan	-0.036***	-0.003
	(0.010)	(0.011)
BPR × British Colony	0.029	
	(0.020)	
Steam Intensity		-0.744**
		(0.297)
Observations	1395	690
R^2	0.234	0.309
Country fixed effects	✓	✓
Sample	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}. BPR stands for "British Patent Relevance", it captures how relevant British patents are to the vocabulary used in manuals of an industry k. BPR is standardized to have a mean of 0 and a standard deviation of 1. 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 H.6 for details about the data construction). Robust standard errors are in parentheses. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

Table A.6: Annualized Export Growth and British Patent Relevance - Manufacturing Sectors

	Export Growth						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
BPR × Japan	0.124**	0.124**	0.124**	0.124**	0.124**	0.124**	0.124**
	(0.055)	(0.055)	(0.055)	(0.055)	(0.055)	(0.055)	(0.055)
BPR × Not Japan		-0.014	-0.015	-0.023	-0.041*		
		(0.012)	(0.013)	(0.014)	(0.022)		
BPR × English-Speaking			0.001				
			(0.025)				
BPR × French-Speaking				0.038			
				(0.023)			
BPR × Top-4 Codified					0.050**		
					(0.024)		
BPR × High-Income						-0.347	-0.347
						(0.335)	(0.335)
BPR × Medium-Income						-0.041	0.079
						(0.885)	(0.911)
BPR × Low-Income						-0.856	-0.318
						(0.760)	(0.967)
BPR × Asia							-1.180
							(1.153)
Observations	31	661	661	661	661	661	661
R^2	0.133	0.362	0.362	0.364	0.366	0.363	0.364
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}. BPR stands for "British Patent Relevance", it captures how relevant British patents are to the vocabulary used in manuals of an industry k. BPR is standardized to have a mean of 0 and a standard deviation of 1. 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.7: Annualized Export Growth and British Patent Relevance: Dropping Regions

	Export Growth, Dropping Exports to						
	(1) English-Speaking	(2) British Colonies	(3) Languages Similar to English	(4) High-Income	(5) Medium-Income	(6) Low-Income	(7) Asian
British Patent Relevance	0.112***	0.112***	0.112***	0.089***	0.111***	0.108***	0.108***
	(0.031)	(0.031)	(0.032)	(0.031)	(0.032)	(0.036)	(0.036)
Observations	71	71	71	70	67	61	61
R^2	0.107	0.107	0.108	0.065	0.107	0.076	0.076
Constant	✓	✓	✓	✓	✓	✓	✓
Sample	Japan	Japan	Japan	Japan	Japan	Japan	Japan

Note: The dependent variable, "Export Growth," is the annualized export growth rate for industry k between {1880,1885} and {1905,1910}. BPR stands for "British Patent Relevance", it captures how relevant British patents are to the vocabulary used in manuals of an industry k. BPR is standardized to have a mean of 0 and a standard deviation of 1. 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. $^{*}p < 0.10$,$^{**}p < 0.05$,$^{***}p < 0.01$.

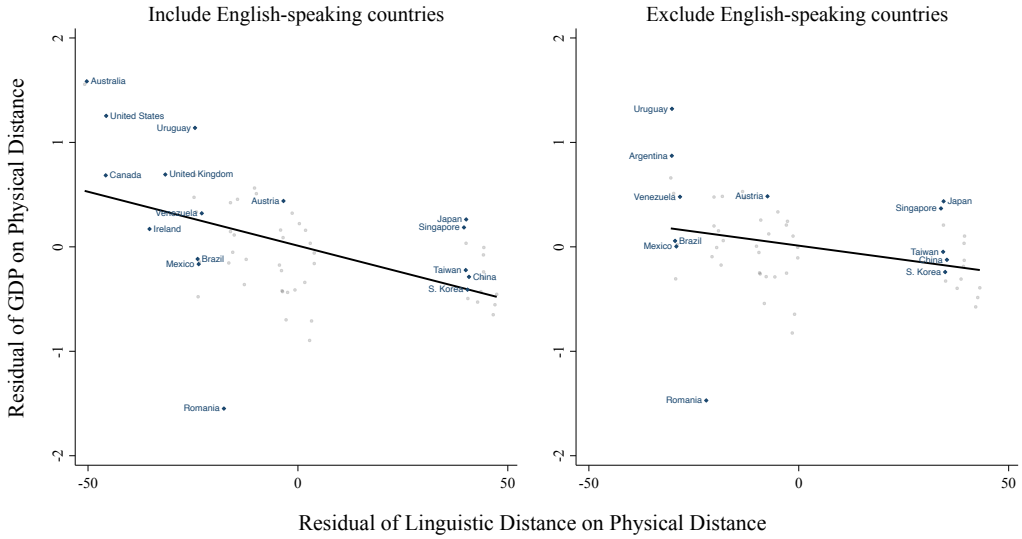
Table A.8: Annualized Export Growth and British Patent Relevance: Dropping Sectors

	Export Growth, Dropping		
	(1) Cotton-Textiles	(2) All Textiles	(3) Iron and Fabricated Metals
British Patent Relevance	0.121***	0.111**	0.123***
	(0.036)	(0.045)	(0.034)
Observations	69	63	69
R^2	0.112	0.086	0.111
Constant	✓	✓	✓
Sample	Japan	Japan	Japan

Note: The dependent variable, "Export Growth," is the annualized export growth rate for industry k between {1880,1885} and {1905,1910}. BPR stands for "British Patent Relevance", it captures how relevant British patents are to the vocabulary used in manuals of an industry k. BPR is standardized to have a mean of 0 and a standard deviation of 1. Each column drops exports to a particular industry or group of industries. (1) drops cotton textile-related industries. (2) drops all industries related to textiles. (3) drops industries related to producing iron. Robust standard errors are in parentheses. $^{*}p < 0.10$,$^{**}p < 0.05$,$^{***}p < 0.01$.

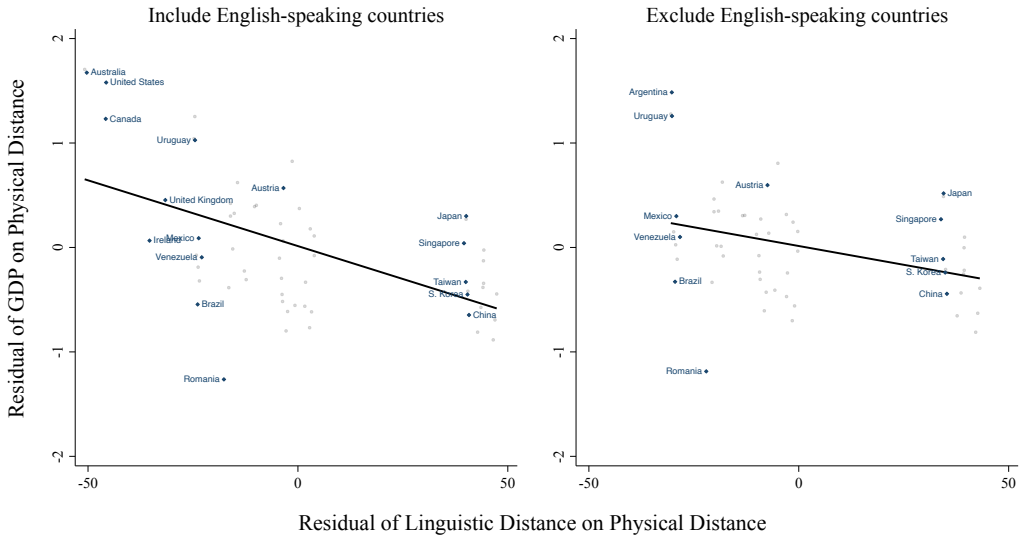
F Additional Figures

Figure A.5: Linguistic Distance Partial Regression Plot for 1870



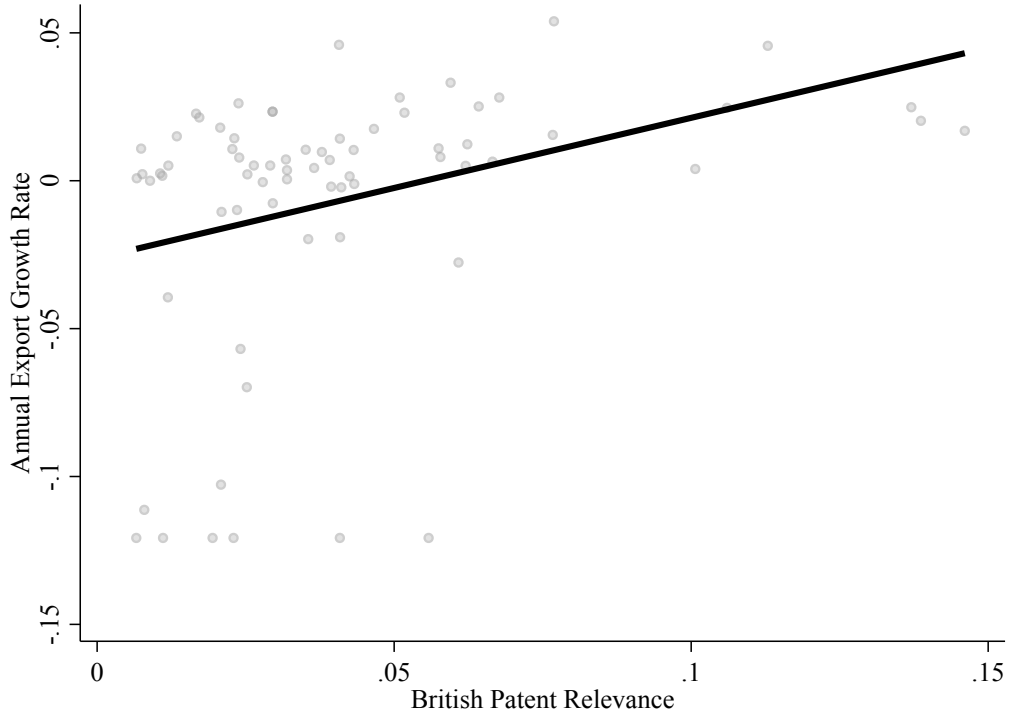
Note: This figure plots the relationship between log GDP per capita in 1870 and linguistic distance after controlling for log physical distance. Data are from the Maddison dataset, the U.S. Department of State's Foreign Service Institute, and *CEPII*, respectively.

Figure A.6: Linguistic Distance Partial Regression Plot for 1913



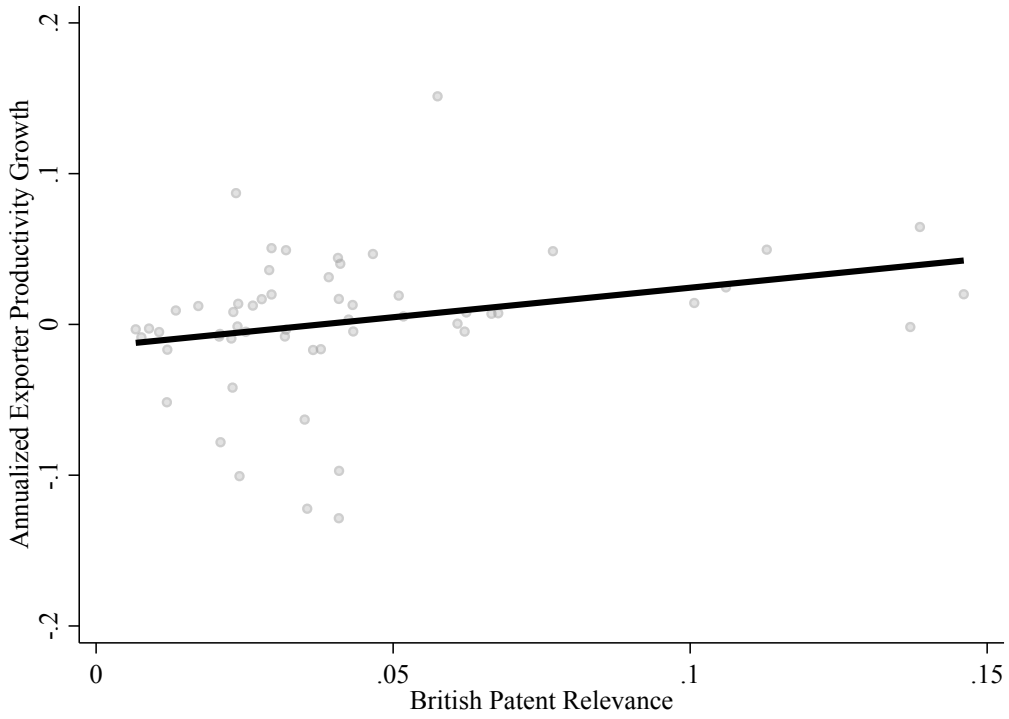
Note: This figure plots the relationship between log GDP per capita in 1913 and linguistic distance after controlling for log physical distance. Data are from the Maddison dataset, the U.S. Department of State's Foreign Service Institute, and *CEPII*, respectively.

Figure A.7: 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}. BPR stands for "British Patent Relevance", it captures how relevant British patents are to the vocabulary used in manuals of an industry k. BPR is standardized to have a mean of 0 and a standard deviation of 1. See text for details on variable construction.

Figure A.8: 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}. BPR stands for "British Patent Relevance", it captures how relevant British patents are to the vocabulary used in manuals of an industry k. BPR is standardized to have a mean of 0 and a standard deviation of 1. See text for details on variable construction.

G Bilateral Trade Dataset

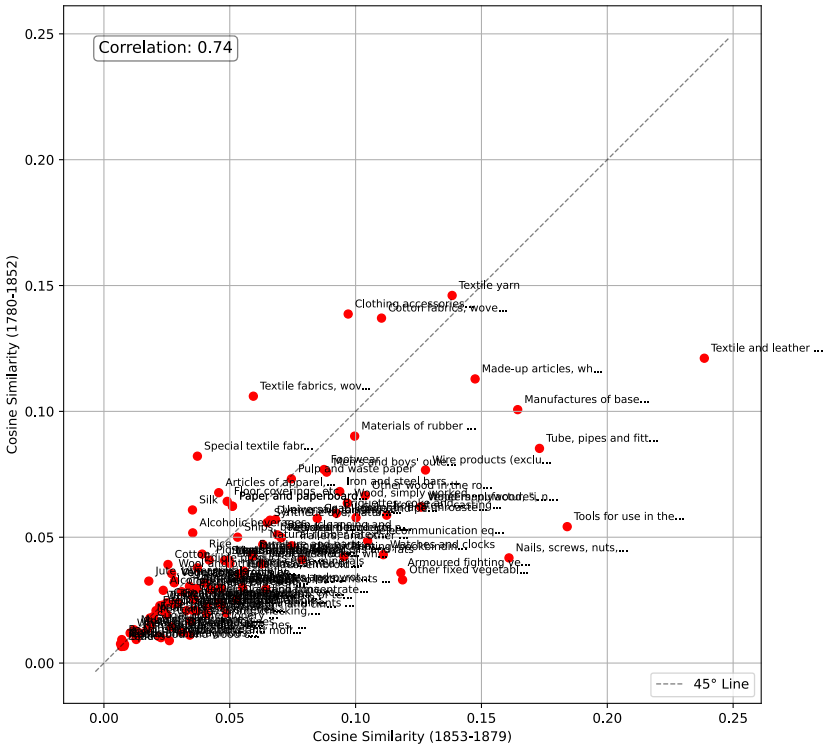
Our master bilateral, product-level trade dataset is constructed from four main sources:

1. **Belgian manufacturing exports and imports in 1880, 1885, 1905, and 1910.** We obtain the Belgian bilateral manufacturing product-level trade data 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 Revision 2 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.

2. **Italian exports to and imports from top trading partners in 1880, 1885, 1905, and 1910.** We obtain Italian trade data from Federico et al. (2011). This dataset harmonizes historical trade records from Italian customs between 1862 and 1950 by reconciling the different product lines to SITC Revision 2 codes. The source reports bilateral trade at the product level between Italy and its ten biggest trading partners.

3. **American exports and imports in 1880, 1885, 1905, and 1910.** 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). We digitized and concorded these data to SITC Revision 2 codes.

Figure A.9: Cosine Similarities between 1780-1852 and 1853-1879



Note: This plot compares cosine similarities constructed using British Patents from 1780-1852 (y-axis) against cosine similarities using British Patents from 1853-1879 (x-axis).

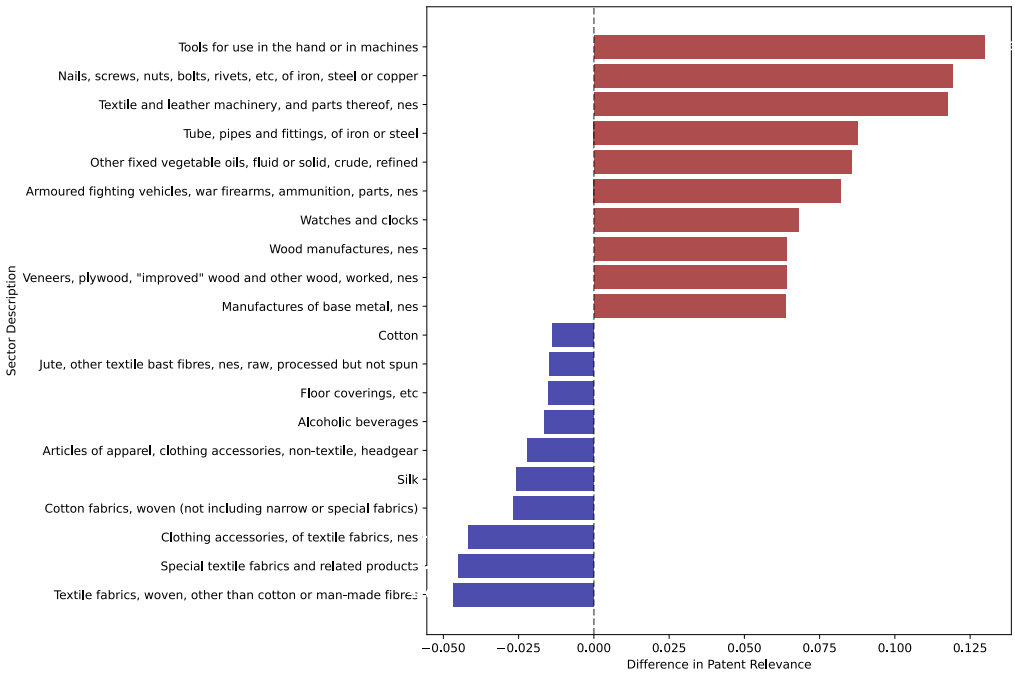
4. **Japanese exports and imports in 1875, 1880, 1885, 1905, and 1910.** We obtained bilateral product-level Japanese export data at five-year intervals between 1880 and 1910 from Meissner and Tang (2018). We digitized and concorded the Japanese export data for 1875. The Japanese trade data were sourced 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 use only the tables from the "Quantity and Value of Commodities Imported/Exported from Various Countries" sections.

Japan and the U.S. kept detailed records of their trade with foreign countries between 1880 and 1910. We used the Meissner and Tang (2018) product-SITC mapping wherever possible for Japan and the U.S. to ensure consistency. Each entry provides the name of the product, its quantity, units, transaction value, and year, as well as the names of the exporting and importing countries. The construction of these data involves digitizing the records and harmonizing products and country names. To construct a harmonized dataset across different reporting countries, we convert all data to a common currency, harmonize country names, and address issues of double reporting. The protocols we adopted are described in detail in the subsections below.

G.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

Figure A.10: Sectors with Highest Positive and Negative Changes in Cosine Similarities



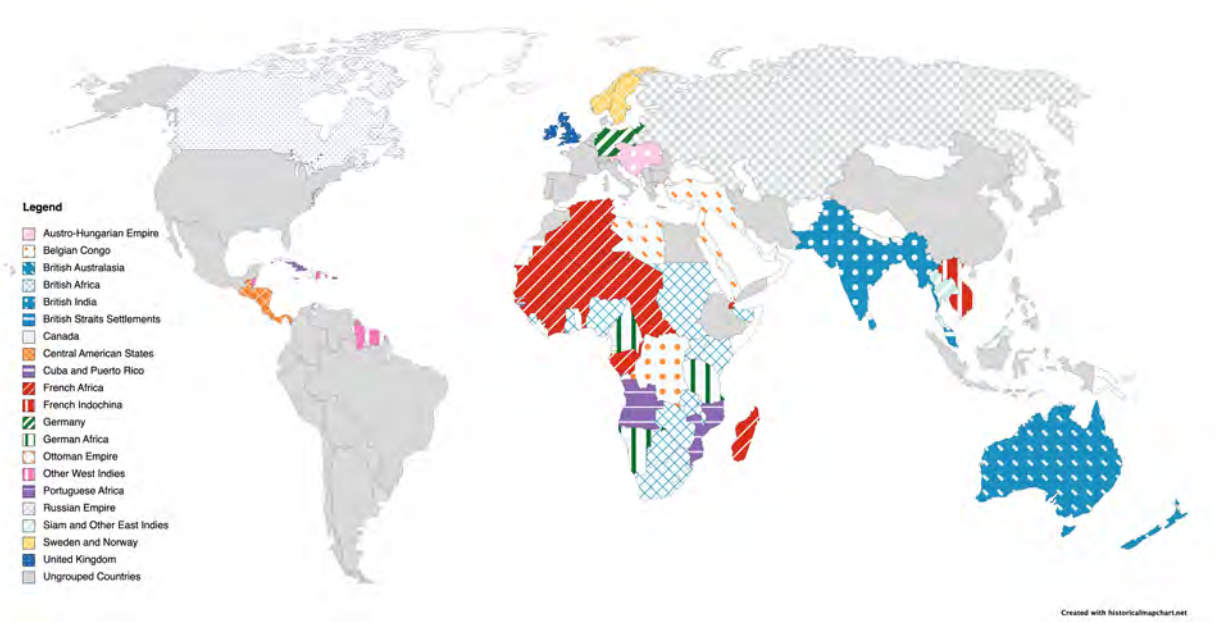
Note: This plot compares sectors with the highest positive and negative changes in cosine similarities based on the 1780-1852 sample and the 1853-1879 sample. A positive change means that the cosine similarity was higher in the 1853-1879 sample than in the 1780-1852 sample.

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 or 1885 books *and* the 1905 or 1910 books for each reporter.
4. If the country group did not meet the previous requirement, then we built a regional group that did. For example, Honduras, Nicaragua, and Costa Rica do not have three reporters in 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 distinct entities, each with substantial trade volumes in our dataset. If one country, in one year, reported “Hong Kong & Singapore," we dropped this observation.

Appendix Figure A.11 illustrates how we grouped countries. We use the map of the world on the eve of World War I (1914) as a baseline for our country groups.

Figure A.11: Country Groups



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, and countries in gray were not modified. 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 is 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.

G.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 did 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.

H Other 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.

H.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)."

H.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.5 - A.6. To do so, we obtain the modern (2023) plurality language spoken in each country from "Ethnologue".

H.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.5 - A.6.

H.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.5 - A.6. To do so, we use data from *Centre d'Etudes Prospectives et d'Informations Internationales* (CEPII) 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.

H.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 (Maddison Project Database, 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.5 - A.6. Further, we also use this source to assign regions into income groups in the main analysis (Section 5).

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 A.2.

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).

H.6 Steam Intensity Usage: Chanut (2000)

In Table A.5, we control for the intensity of steam usage of industries in our regressions. We measure this variable based on 19th-century 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, where the wage bill is defined as:

$$\begin{aligned}\text{Wage Bill} = \quad & \#\text{ of Male Workers} \times \text{Avg. Male Hourly Wage} \quad + \\ & \#\text{ of Female Workers} \times \text{Avg. Female Hourly Wage} \quad + \\ & \#\text{ of Child Workers} \times \text{Avg. Child Hourly Wage}\end{aligned}$$

H.7 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.

I Constructing the British Patent Relevance measure

I.1 Overview

In our empirical analysis, we develop a method to quantify the supply of codified knowledge generated by the IR for each industry. We use a textual approach that follows how codified technical knowledge was disseminated in this period: through the publication of technical manuals. For each industry, we measure the textual similarity from historical technical manuals (in English) and patents. We call this measure British Patent Relevance (BPR). We also construct an analogous measure using U.S. patents, which we call U.S. Patent Relevance (USPR) measure. To implement this, we assign at least one technical manual describing production techniques to each SITC industry code and compute the similarity of its text to either British or U.S. patent texts.

We construct unigrams (e.g., *steam*) and bigrams (e.g., *steam engine*) from both patent text and technical manuals. These terms are stemmed (e.g., *steam engine* → *steam engin*) and aggregated into an industry-level corpus, with one corpus for each industry k. Patent text forms a separate corpus. For each corpus, we compute a TF-IDF (Term Frequency-Inverse Document Frequency) vector that characterizes its vocabulary. Patent relevance for industry k is then defined as the cosine similarity between the TF-IDF vector of industry k's technical manuals and that of the patent corpus. We describe each step in detail below.

I.2 Building the Terms

We construct terms from the raw text by generating n-grams. The procedure is as follows:

1. Split the raw text into sentences.
2. Convert all words to lowercase, stem them, and standardize spelling (UK spelling → US spelling).
3. Represent each sentence as an ordered list of words.
4. Generate n-grams from each sentence word list.
5. Count the frequency of each n-gram within a sentence and aggregate across sentences.
6. Remove n-grams that contain at least one stop word (e.g., "a," "the").
7. Produce a dataset containing all n-grams in the document and their corpus-level frequencies.

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

I.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.

I.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:

$$\text{TF}(\tau, c) \equiv \frac{F_{\tau,c}}{\sum_{\tau' \in c} F_{\tau',c}} \tag{A.13}$$

where $F_{\tau,c}$ is the raw count of term τ in corpus 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 the corpus) as:

$$\text{IDF}(\tau, C) = \log\left(\frac{N}{N_\tau + 1}\right) \tag{A.14}$$

where N is the total number of documents (books and patent[9]) in C; N_τ is the number of books in the corpora where the term τ appears.

The TF-IDF is then

$$\text{TF-IDF}(\tau, c, C) = \text{TF}(\tau, c) \cdot \text{IDF}(\tau, C) \tag{A.15}$$

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 Patent Relevance of industry k as the similarity between the TF–IDF vector of its technical manuals and the TF–IDF vector of patent texts. The intuition is that if industry manuals use vocabulary similar to that found in patents, then patents contain knowledge relevant to that industry. We measure similarity using cosine similarity, the standard NLP metric for comparing text representations.

Cosine similarity corresponds to the cosine of the angle between two vectors. In the case of our baseline results, it compares the vector of word frequencies in the Bennett Woodcroft patent collection (British patents), $\boldsymbol{BW}$, with the vector of word frequencies in the technical manuals for industry i, $\boldsymbol{TM_i}$. Formally,

$$BPR_i \equiv \frac{\boldsymbol{BW} \cdot \boldsymbol{TM_i}}{\|\boldsymbol{BW}\|\|\boldsymbol{TM_i}\|} = \frac{\sum_{j=1}^{n} BW_j TM_{ij}}{\sqrt{\sum_{j=1}^{n} BW_j^2}\sqrt{\sum_{j=1}^{n} TM_{ij}^2}}, \tag{A.16}$$

where BPR_i denotes the *British Patent Relevance* of industry i. By construction, BPR_i lies between 0 and 1. A value of 1 indicates that industry manuals and patents use exactly the same vocabulary in the same proportions, while a value of 0 indicates no overlap in vocabulary.

I.5 Data Sources

Industry For each industry k (defined by SITC-3 Revision 2 codes), we hand-curated a list of nineteenth-century books describing the production process of the goods produced by k from

[9]The whole set of patents counts as one document.

HathiTrust. We picked the technical books that best matched the knowledge an entrepreneur would have had access to if they had studied Western knowledge before Japan began to industrialize, i.e., before the 1880s.

British Patents (1617-1852): The patent text from British patents between 1617-1852 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). These documents contain a synopsis of each patent published between 1617 and 1852. The document is divided by categories, where each patent can be categorized into one or more categories. We digitize the text of these documents and drop duplicated patents (i.e., patents that are in more than one category). Our baseline analysis uses only patents published between 1780-1852. This data was obtained through *HathiTrust*.

British Patents (1853-1899): For this period, we rely on the digitized collection of British patents compiled by Coluccia and Dossi (2025). Their data contains the full text of all British patents published between 1853 and 1899. We treat this period separately from 1617-1852 because of the major patent reform of 1852, which reduced filing costs by roughly 75% and triggered a fivefold increase in patenting within a single year. Moreover, while the pre-1852 data consists only of short synopses, the 1853-1899 dataset provides full patent descriptions. To avoid concerns about the comparison between full patent descriptions and patent synopses, we present a version of BPR (1853-1879) summarizing the full patent descriptions so that they have a similar length to patent synopses. To do this, we used OpenAI's API with the following prompt:

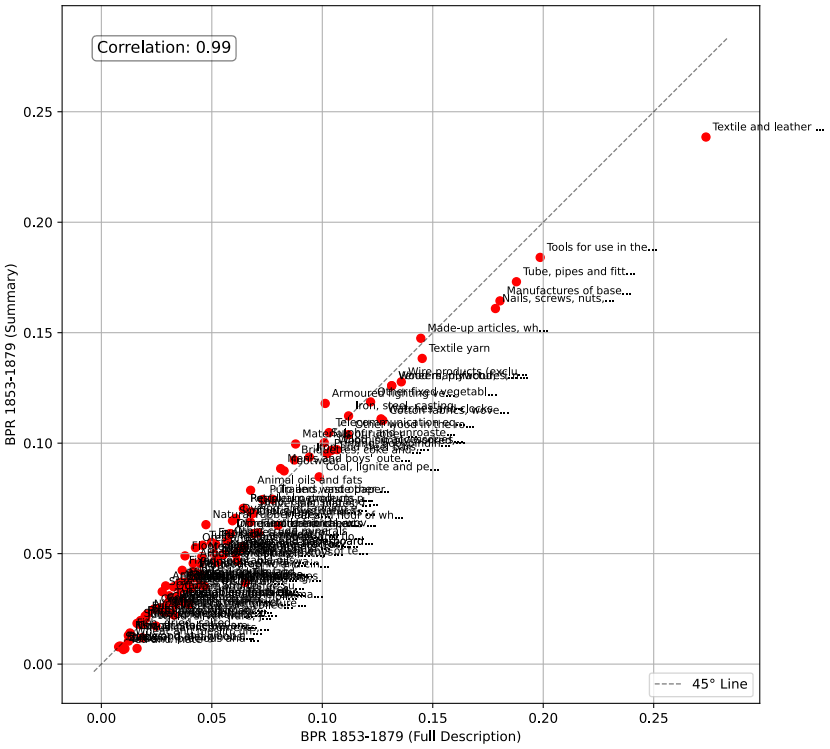
> *Summarize the following 19th-century British patent in MAX. 15 words. Focus strictly on the technical content, state what the invention is, and describe the mechanism or process. Use only vocabulary found in the patent itself or in common use at the time of application. Omit the author and date from the summary. Do not start with phrases like 'This invention describes'.*

The 15-word limit mirrors the average length of the synopses between 1617 and 1852, ensuring comparability across periods. We construct BPR for the 1853-1879 period (right before our analysis with trade data starts). To address potential concerns about the process of summarizing the patent descriptions, we also computed our BPR measure for the 1853-1879 period without summarizing the patent descriptions. The cosine similarities using full descriptions or summaries are very similar, with a correlation of 0.99, as can be observed in Figure A.12.

U.S. Patents (1836-1910): We collect U.S. patent descriptions from 1836 (the earliest year available) through 1910 by web-scraping *Google Patents*, which provides digitized versions of all U.S. patents. Our scraper builds on the tool developed by Kelly et al. (2021) and extracts the patent number, title, date, and full description for each patent.

U.S. patent descriptions typically begin with formulaic phrases such as "To all whom it may concern, be it known that (. . .)". We identify the most common introductory phrases and remove them so that descriptions begin directly with the technical content. Google Patents digitization relies on Optical Character Recognition (OCR), which can introduce transcription errors. To mitigate this problem, we retain only words appearing in the Oxford English Dictionary (which has over 500,000 entries). Words that are not in the dictionary are treated as OCR errors and discarded. On average, this cleaning step removes about 3% of words in a typical patent description.

Figure A.12: BPR (1853-1879) Computed with Full Descriptions and Summaries



J 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.[10] The number of new words by year can be seen on Figure 3.

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

K.1 Overview

We report the source libraries for our data on technical books in Table A.9. 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

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

(possibly published in different years) with the same book ID. Importantly, the number of books 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.[11] 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.

[11]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.9: 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	1500-2023	Nippon Decimal Classification	Technology (500) Industry (600)
Korean National Library	Link	Korean	1500-2023	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	1500-2023	Chinese Library Classification System	Agriculture (S) Industry (T) Transportation (U)
National Central Library (Taiwan)	Link	Chinese	1500-2023	Keyword Search	Hand-constructed
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

K.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.