

Publishing Nations: Technology Acquisition and Language Standardization for European Ethnic Groups

YU SASAKI

This article examines the causes and variation of language standardization across European ethnic groups from a historical perspective. Although language has long garnered interest in the study of ethnicity and nationalism, how language becomes standardized has yet to be offered. In this article, I argue that the acquisition of the printing press is critical to explaining the occurrence and variation of standardization. Using the first publication of vernacular dictionaries as a proxy for standardization, I present a systematic investigation of the standardization process for 171 ethnic groups in Europe from 1400–2000 CE. Empirical tests come from an original data set that collects information on political, economic, and social dimensions. Findings from event history models show that (1) printing press adoption is positively and significantly correlated with vernacular dictionaries; and (2) early adopters are more likely to standardize vernaculars than latecomers.

The causal connection between culture and economic growth has garnered increasing attention. Although the broad linkage is traced back to Adam Smith, recent work focuses on specific functions of culture. Some, for instance, argue that wealth may in part be driven by an “inclusive” culture in society, where no one is called heretical when challenging the conventional wisdom, or where there is an institutionalized medium, such as journals, that guarantees the freedom of speech for ideas that challenge orthodoxy (Mokyr 2002, 2016; McCloskey 2016). Empirical analysis has demonstrated that certain individual beliefs and preferences, when geographically concentrated, are growth enhancing (Guiso, Sapienza, and Zingales 2006; Tabellini 2008, 2010). Others show that inclusivity matters but excessive diversity may have adverse effects by incurring high transaction costs (Ashraf and Galor 2013).

The Journal of Economic History, Vol. 77, No. 4 (December 2017). © The Economic History Association. All rights reserved. doi: 10.1017/S0022050717000821

Yu Sasaki is a Project Researcher, Institute for Advanced Studies on Asia, University of Tokyo, 7-3-1 Hongo, Bunkyo-Ku, Tokyo, 113-0033, Japan. E-mail: ys253@uw.edu.

I thank Editor Ann Carlos and two anonymous reviewers for constructive feedback. I also thank Jeffrey Arnold, Aaron Erlich, Anthony Gill, Darryl Holman, Edgar Kiser, James Long, David Lopez, Victor Menaldo, Steven Pfaff, Jared Rubin, and participants of the Severyns-Ravenholt Seminar in Comparative Politics at the University of Washington, October 2014, for helpful comments. All errors are mine.

Among the myriad attributes of culture, language offers a tangible and testable causal channel to economic growth. Language, in theory, can be conceptualized as a medium to gain market access, information, public goods, and rents. Many societies use language as a political instrument to build barriers to entry or deny certain groups access to market transactions and economic resources (Laitin and Ramachandran 2016; Liu 2014). Language can also function as a focal point around which to build effective political coalitions (Fearon 1999). This is in part why speakers of the same language in highly diverse societies typically coalesce as an ethnic group in the competition for political power and economic opportunities.¹ Thus, the ability of ethnic groups to use language as an instrument of power can have profound consequences, not only on the economy, but for society as a whole.

Few theoretical and empirical works systematically address why some ethnic groups gain the ability to use language more effectively than others. Although social science research has offered general hypotheses, the roots of the standardization of core cultural attributes for ethnic groups, especially language, have not been widely investigated.² It is one of the most difficult cultural attributes to standardize, because, compared to flags and anthems, *codified* language requires literacy.³ Benedict Anderson (2006) comes closest to theorizing about language rationalization. His thesis connects the profit motive of printers and booksellers (“print-capitalism”), which first arose following the invention of the printing press in late fifteenth-century Europe, to “national consciousness” (Anderson 2006, Chapter 3). Yet, as I show later, a broad description of this linkage leaves unaccounted multiple possible causal mechanisms. Moreover, systematic evidence that tests Anderson’s hypotheses about language standardization has not yet been offered.

I fill these gaps by providing a simple conceptual framework and empirical analysis of how language standardization occurs for European ethnic groups. I distinguish two plausible channels through which Anderson’s hypothesis that the acquisition of the printing press could spur the greater use of the vernacular over Latin. The first is the selection process. Here access to print technology gave early-modern ethnic groups an impetus to make innovations, including in later periods vernacular codification when such development was deemed necessary and viable. Accessibility

¹ For empirical works on the relationship between ethnic diversity and economic performance, see Alesina, Devleeschauwer, Easterly, et al. (2003), Alesina and La Ferrara (2005), Desmet, Ortuño-Ortín, and Wacziarg (2012), Easterly and Levine (1997).

² For major works, see Gellner (2006) on the impact of industrialization, Anderson (2006) on the role of literacy and capitalism, Smith (1986) on the availability of ancient symbols.

³ In this article, I use “language standardization,” “language rationalization,” and “language codification” interchangeably.

is critical because fixing language was never a reason for the spread of presses. Ethnic groups with a sovereign state benefited from this channel. The second process is the one whereby standardization was a conscious choice on the part of early-modern printers to promote vernacular publications to meet their interest in education, proselytization, and profit. Many early-modern ethnic groups were able to codify their tongue in modern times without independent statehood by taking advantage of this coincidence. Multiple routes to the equifinality between print technology and language standardization reflect the wide-ranging effects of the printing press (Eisenstein 1979; Dittmar 2011).

This article presents a new data set to explore these links with statistical evidence for 171 European ethnic groups for the period between 1400 and 2000 CE. The evidence is consistent with the hypothesis that the acquisition of the printing press is positively associated with standardization of the vernacular and that early adoption of that technology corresponds to greater chances of standardization. Using event history models, I find that ethnic groups that adopted the press have a four to nine times increase in the chance that their vernacular gets standardized. Similarly, the evidence shows that early adopters were six percentage points more likely to rationalize their vernacular in the modern period than were latecomers. The results hold whether or not ethnic groups comprised an independent state. The main findings are based on Cox proportional hazard models. But I also test their robustness using logistic regression models. I address endogeneity concerns, first, by accounting for the possibility that the spread of print technology is driven by human capital and, second, by using the geographical distance to Mainz as a source of exogenous variation for the spread of presses. All models confirm that the adoption of print technology has a substantive and significant impact on language standardization.

In this article, I make two contributions. First, I seek to demonstrate that there are non-state routes accounting for the variation in cultural consolidation among ethnic groups. Territorial sovereignty plays so integral a role in cultural preservation that it is commonly embraced in the literature on ethnicity and nationalism as a precondition for the construction of nations (Anderson 2006; Breuilly 1993; Gellner 2006; Hechter 2000). Yet recent research on the causes of ethnolinguistic diversity has empirically shown that the state is not the only determinant and that it is critical to consider non-state determinants such as geography and technology (Ahlerup and Olsson 2012; Ashraf and Galor 2013; Laitin, Moortgat, and Robinson 2012; Michalopoulos 2012; Spolaore and Wacziarg 2009). This article focuses on the impact of access to print technology and the timing of technological access to address variation in a core attribute of ethnicity, namely, language standardization. My second contribution is to show

deep historical roots of contemporary politics that harnesses ethnicity and culture as political instruments. An emerging body of empirical work in economic and political history demonstrates that the impact of historical events can have centuries-long persistence on current economic and political outcomes (Comin, Easterly, and Gong 2010; Nunn 2014; Olsson and Paik 2016; Spolaore and Wacziarg 2013; Stasavage 2014). This article joins this growing literature by providing evidence that contemporary ethnic groups' ability to use language for political participation and access to the market was determined many centuries ago.

CULTURAL RATIONALIZATION, ETHNICITY, AND THE NATION

The unit of analysis in this article is the ethnic group. Before laying out hypotheses, it is critical to define "ethnicity" and the related concept, the "nation," and address the relationship between the two. I define ethnicity as a category that delimits membership to a group and ethnic group as a collectivity determined by a set of attributes shared among its members. Following many scholars, I use the myth of common descent as the most basic characteristic for the definition.⁴ As explained later, many ethnic groups attempted to consolidate their cultural practices as a path to political survival in a centuries-long competition over resources and territory, which was particularly intense in Europe (McNeill 1982).⁵ Any symbols reminiscent of common origin may be selected for group consolidation, such as a memorable landscape, historic buildings, clothing, or a flag. By contrast, the nation may be defined as a collectivity with standardized culture that commands loyalty among the community members. Standardization here is synonymous with institutionalization;⁶ a nation often designates certain cultural practices to be written as "official," as in an anthem, interpretations of its origins, and language use, which are consistently invoked to consolidate group membership. Successful language standardization, in particular, requires literacy among the members of the nation so that official culture is effectively promulgated and sustained from one generation to the next. Ethnicity, in short, is distinguished from the nation in terms of standardization.

When ethnicity and nation are understood in this way, a few theoretical and empirical implications emerge. First, the ethnic group can be

⁴ Weber (1978 [1922]) is oft-invoked. An extensive review of the definition of ethnicity and its implication for political analysis is found in Chandra (2006).

⁵ Tilly (1975, 1992) and Ertman (1997) are also associated with this interpretation, using the state as unit of analysis.

⁶ That standardization is synonymous with institutionalization is consistent with the definition of standardization in linguistics. See Brown (2006b, p. 121).

an analytical building block when considering questions about the nation. This approach is useful not only for conceptual distinction but also for empirical analysis. Second, it renders the category—nation—more amenable to concrete observation, which opens a way for causal identification. One such strategy may be to identify the timing of standardization as one salient dimension of ethnic groups' cultural attributes and examine the mechanism by which standardization occurs.

Among the multitude of cultural practices, the standardization of the vernacular is one of the most important in specifying the mechanism through which ethnic groups become consolidated nations. The rationalization of non-linguistic dimensions, say an anthem, may precede that of linguistic ones, but enforcement would predictably be challenging in the absence of a shared means of communication. By contrast, once a vernacular is codified, this would drastically lower the cost of access to information, thereby making consolidation of the other dimensions of culture more efficient. In a classic study of cultural rationalization in Third-Republic France in the late nineteenth century, Eugen Weber (1976, p. 313) recounts the centrality of standardized language instruction at public school in the countryside. Weber vividly illustrates how a pupil who uttered words in an “unauthorized” tongue—*patois*—was chastised by having to hold a display showing a *faux pas* until the next student committed one. Further, a unitary language has a spillover effect in social organization. For instance, it makes the administration of military-related tasks such as recruitment, training, and the solidarity of personnel effective and therefore enhances fighting capacity. Uniform language fulfills a variety of functions, not just as a chief repository of cultural knowledge, but also as a useful instrument for policy.

My focus on language is certainly not new. In fact, language plays a central analytic role in the literature on ethnicity and nationalism (Anderson 2006; Bell 2001; Gellner 2006; Kohn 2005). Convention holds that a nation may be observed when a government is conducted in the predominant tongue of its territory (Gellner 2006; Hechter 2000; Hobsbawm 1990; Tilly 1994). The extant literature indicates that fixing the vernacular is assumed to be critical to achieve what Andreas Wimmer calls “the rule of like over like” (2002, p. 213). Across Europe, language historically served as one of the most important cultural dimensions with which to define membership to the given nation. The rise of modern citizenship, an institution in which language fluency is typically a requirement, is a good illustration.⁷ Indeed, contemporary works use language

⁷ Brubaker (1992) is most closely associated with this view.

as the dependent variable to theorize about the variation in the salience of identity across time or within an ethnic group (Laitin 1998, 2007; Brubaker 2004) and the variation of “language regimes”—state policy on language instruction (Albaugh 2015; Cardinal and Sonntag 2015; Liu 2014; Safran and Laponce 2005).

A systematic investigation as to why language rationalization has become the default choice for ethnic groups seeking survival has not yet been undertaken. There is little doubt that the codification of the vernacular is, in theory, a highly costly project. It entails the classification of speech into main “trunks” and “branches” (as in the language trees), the transcription of speech into letters, the establishment of a grammar and related rules on usage, and finally the construction of an orthography, a set of rules about spelling. Once these rules are clearly written out and consistently used and taught among the speakers, a language may be said to have been “standardized.” Publication of a dictionary may come at the very end of this lengthy process. Competent and devoted experts would be needed to bring the project to completion. It is, therefore, easy to imagine that each step is a labor-intensive, time-consuming, and financially demanding process. High fixed costs should make it clear that language standardization is not a “natural” choice. Yet, empirically, no ethnic groups that are well-known, such as the English, French, Italians, and Russians, fail to have their languages codified. Lesser-known ones, aspiring to survive, attempt to follow suit by consciously using their tongues in everyday communication and school instruction.⁸ Two questions arise with respect to high fixed costs in labor and time on language rationalization. The first is a theoretical question. Why do some ethnic groups make such a costly investment and fix their vernacular? The second is empirical. Do hypotheses on the first question explain the variation of language standardization across ethnic groups? Addressing these questions fills an important lacuna in the study of ethnicity, nations, and ethnic diversity.

PRINT TECHNOLOGY AND LANGUAGE STANDARDIZATION

The Technological Innovation

I hypothesize a positive relationship between the printing press and language standardization. But it is crucial, first, to consider the technology’s broad impact beyond language. The most general and profound

⁸ Minahan (2000) and Price (1998) offer an abundance of examples, such as Welsh and Frisian, about the active use of the vernaculars by ethnic groups that constitute a demographic minority in their state of residence.

effect that print technology brought to bear would be to reduce the cost of access to information (Mokyr 2005; Bernstein 2013). Metal movable typography, which was invented by Johannes Gutenberg, among others, circa 1450 in Mainz, Germany, greatly encouraged innovations in knowledge production on many scales. Innovations occurred more cheaply, spread more widely and more quickly, and the incentives to produce grew stronger.⁹ The diffusion of printing technology was remarkably swift by fifteenth-century standards. More than 110 cities had a press established by 1480 (Febvre and Martin 1976, p. 182) and the number grew to over 240 by 1500 (Clair 1976). Thereafter, book production surged dramatically. There were an estimated 5 million manuscripts produced during the fifteenth century in a dozen European countries, a 358-fold increase from the sixth century and an 82 percent increase from fourteenth-century production levels (Buringh and van Zanden 2009, p. 416). Similarly, printed book production increased 6.3 times in the first half of the sixteenth century from the 12 million books printed during the *incunabula* period of 1450–1500 (Buringh and van Zanden 2009, p. 417).¹⁰ In this period, the price of the book dropped by two-thirds (Dittmar 2011). Moreover, the technology significantly reduced person-hours to the extent that it pushed scribes out of business (Bernstein 2013). Figure 1 documents the rapid diffusion of printing technology among the European ethnic groups in my data set (described in detail later). Of the 96 groups that acquired the press, approximately half (41.7 percent) did so during the *incunabula* period. Figure 1 counts the first year of print adoption for each group.

The State's Role in the Spread of Print and Language Standardization

The reduced cost of access to print media served as a catalyst for activities beyond book production.¹¹ The initial demand for the new technology came primarily from university instructors, clergy, lawyers, and wealthy merchants (Febvre and Martin 1976, pp. 172–80). Yet the cost

⁹ Mokyr (2005, pp. 1123–24) calls this phenomenon “propositional knowledge,” the type of knowledge with which to describe, characterize, and theorize—that is, innovate. Examples are identified in Footnote 11. This is contrasted to “prescriptive knowledge,” the type of knowledge to produce or invent. Vernacular dictionaries fit this category. The movable-type printing press is a general-purpose technology that produces both types of knowledge.

¹⁰ *Incunabula* refers to the first 50 years of printed book production after the Gutenberg press.

¹¹ Historical and empirical research shows that the printing press spurred commerce and economic growth (Dittmar 2011; McCusker 2005); literacy (Eisenstein 1979; Graff 1987; Buringh and van Zanden 2009); the Protestant Reformation (Rubin 2014); and democracy outside of Europe (Woodberry 2012).

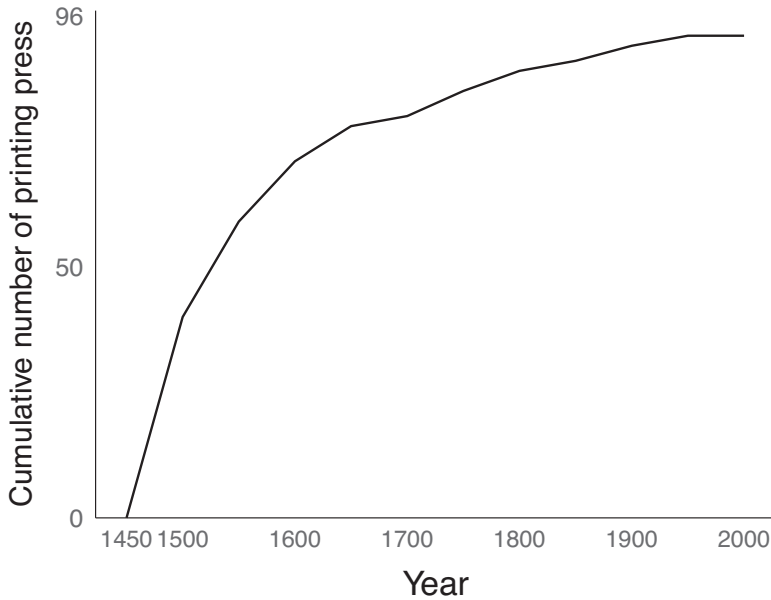


FIGURE 1
CUMULATIVE NUMBER OF PRINTING PRESS ADOPTION AMONG 96 EUROPEAN
ETHNIC GROUPS

Source: Clair (1976), Febvre and Martin (1976), among others.

of operation remained high and printers were chronically short of capital (Febvre and Martin 1976, Chapter 4). As a result, printers had to work in as many markets, because uncertainty about the emerging market on print media made it a highly risky venture (Pettegree 2010, Chapter 3). For printers, finding a profitable market and sustaining it were equally critical.

States played little initial role in generating supply or demand for the new technology. In premodern Europe, industrial technology like the movable-type press spread primarily through skilled workers rather than states or related political channels (Cipolla 1972). Nevertheless, print technology had a ripple effect in politics by making the dissemination and enforcement of rules cheaper. It proved useful for legal, legislative, and administrative purposes (Graff 1987, p. 109). There is some evidence that states took advantage of print media for two of their core functions: war and taxation. For instance, in the 1510s Maximilian I of the Holy Roman Empire issued propaganda broadsheets and pamphlets to raise manpower and revenue (Pettegree 2010, p. 132). The state also had the potential to induce supply of print by serving as a reliable patron for printers who always looked for one. At the same time, premodern

European states chronically suffered from an unreliable flow of revenue due to a limited degree of institutional centralization (Dincecco 2015). All these rationales point to a favorable condition under which the state could employ printers as royal servants and create a win-win relationship which could be long-lasting, because revenue generation and law enforcement were permanent features of governance. Although printers called for monopoly rights to create a more secure business environment, premodern European states were generally unable to enforce regulations due to their limited institutional capacity (Pettegree 2010, p. 73). Evidence on the political use of the printing press by the state seems indirect at best for the premodern era.

Equally important, independent statehood is not a prerequisite for language standardization. The development of the vernacular for the Bretons offers an illustrative case. Prior to incorporation into France in 1532, there was no clear evidence of Breton being used for official purposes during the five centuries of the independent kingdom (Price 1998, p. 36). However, from around 1450 the Breton language was substantially recorded in manuscripts, and the arrival of the printing press in Rennes in 1485 facilitated stability in language use. The reduced cost of printing led to the publication of a trilingual Breton-French-Latin dictionary in 1499 (the first printed book in Breton) and to the introduction of simplified spelling rules (orthography) in the mid-seventeenth century (Price 1998, pp. 36–37). A greater supply of language use began in the early nineteenth century, when literates produced poetry, history, and novels in Breton, following the publication of grammar books and French-Breton dictionaries in the “purified” form by influential linguist Jean-François Le Gonidec (Hardie 1948, p. 10; Minahan 2000, p. 131). The separatist movement of the twentieth century, though failed, gave an impetus for further linguistic sophistication including spelling unification across the Breton-speaking region of France. Despite the legal exclusion of Breton use in the French school system for 1880–1951, Breton use became more widespread in journals and periodicals, culminating in the 1958 monolingual dictionary, *Geriadur istorel ar brezhoneg (Historical Dictionary of Breton)* (Dalby 1998, p. 64). The example of the Breton language demonstrates, first, that printing technology played an important role in stabilizing the vernacular and, second, that standardization is feasible without an independent polity.

Evidence from the cultural history of early-modern Europe indicates that the spread of the movable type and the development of print media relied less on state actors than private ones. Printers had to secure access to capital other than in the public route, by creating a joint venture, locating

private patrons, and relocating to high-demand cities like Wittenberg, the home of the Protestant Reformation (Pettegree 2010). Thus, the theoretical linkage between the printing press and language standardization is likely to be as strong or even stronger outside the state.

Print and Vernacular Standardization

To begin, I discuss two ways in which the press gave rise to the greater use of the vernacular over Latin. First, printers played a role in triggering competition between vernaculars and Latin. As mentioned in the previous section, print was a high-risk business so that printers had to explore demand in academic and private markets. As Anderson famously points out, after quickly filling the “thin” Latin market, these printers cultivated a “thicker” vernacular one in the lay public, which could only understand the vernacular as no ethnic groups had Latin as a native language (Anderson 2006, p. 38). Indeed, the growing popularity in the use of the vernacular begot the “esotericization of Latin” (Anderson 2006, p. 42), which was made apparent after 1530. The competition largely ended within 80 years following the invention of the Gutenberg press (Febvre and Martin 1976, p. 320).¹²

Second, access to print had an effect of stabilizing languages (McKitterick 1998, p. 296). This point is often overlooked or made in passing in the literature but is crucial. Once words are produced in printed form and widely circulated, they tend to acquire “staying power” in terms of consistency in usage. Furthermore, the technology of mass production enhanced scalability. One conspicuous dimension on which the printing press contributed to language rationalization is spelling. Prior to the invention of the press, spelling was primarily phonetic; thereafter, it became increasingly consistent as the transmission of words was based on a mechanical print medium (Steinberg 1974, p. 125). When printers processed text for publication, they simplified spelling at their own discretion to make their work more efficient (Eisenstein 1979, p. 87). Once types were set, subsequent printers would keep using them when printing the same words to save time. The power of the press to fix

¹² It is important not to attribute the decline of Latin solely to the printing press. As Ostler states, the press occasioned the “second death” of Latin. The “first death” occurred in the late eighth and ninth centuries during which Latin became less frequently used as the means of oral communication in favor of vernaculars (Ostler 2005, Chapter 8). In addition, Latin was not completely displaced and continued to be used over the next few centuries. Febvre and Martin note that “the final blow was struck at Latin” around 1630 due chiefly to the decline in book trade, but that Latin was not eliminated until the beginning of the eighteenth century (Febvre and Martin 1976, p. 330).

language use attracted the attention of school teachers and priests who were concerned about what they deemed as the inconsistent or improper use of language and some of them subsequently became printers themselves. It is crucial to recognize that the invention of the technology did *not* intend to enable or facilitate the rise of the vernacular; standardization was an unintended consequence of it.

What processes then account for the link between the greater use of the vernacular and codification? In *Imagined Communities*, Anderson made an argument about the link between the profit motive of printers and booksellers (print-capitalism) and national consciousness. He theorizes that “print-capitalism gave a new fixity to language, which ... helped build that image of antiquity so central to the subjective idea of the nation” (Anderson 2006, p. 44). This hypothesis describes the processes of language rationalization in general, but two points remain unaddressed. The first is to specify the causal mechanisms linking the press to language standardization. The second highlights the dimension of duration in the standardization process that is highly time-consuming and labor-intensive. The print-capitalism thesis holds that although the vernaculars won the competition with Latin relatively quickly, it does not consider the new race between vernaculars. The new rivalry lasts significantly longer than that the competition with Latin. Latecomers in the game may be severely disadvantaged in that they miss the opportunity to standardize their tongue and are compelled to adopt similar ones that are better codified. These two points merit further discussion.

There are two plausible causal mechanisms from print technology to vernacular codification. The first is the selection process. If early-modern printers established presses in homeland cities of some ethnic groups and started to produce books in the vernacular, the technology would likely prove to be useful later when language standardization became viable or necessary. Even though the original aim of print adoption was printers’ incentive to generate revenue, accessibility to the relevant technology was a catalyst for subsequent innovations. The state is clearly a beneficiary in this hypothesis. Although European states did not create a strong demand for or supply of print, they benefited from the reduced cost of producing print media to facilitate the enforcement of rules and revenue generation. When states’ investment in language rationalization is rewarded by the greater degree of fiscal and institutional centralization, other states may quickly follow suit to stay competitive. Similarly, the selection mechanism can also explain why the language of demographically small ethnic groups was codified. As illustrated in the case of the Bretons, if there was a history of vernacular print, it could allow language entrepreneurs to

revive or solidify their language to achieve national unity and to increase the chances of group survival without an independent state. Thus, the selection process can account for competitive and niche mechanisms.

The second process is conscious choice. Actors may take advantage of the access to printing technology using the vernaculars to advance their interests. Two paths merit attention. The first is the profit motive on the part of the printers. An increase in vernacular publications over those of Latin stirred new demand for translations. The translation business began to flourish in the early sixteenth century with many print offices becoming workshops for translators busily producing classical works in the vernaculars (Febvre and Martin 1976, p. 272). Ethnic entrepreneurs who want to advance learning in their native tongues, such as school teachers and university professors alike, would have an incentive to exploit the declining cost of becoming literate and producing books. Another path is Bible translation. Protestants had an incentive to translate the Latin Bible into the local languages in the proselytization effort. The motive was particularly strong among Protestant reformers. They wrote books and translated Luther's catechism and other writings into many languages. These were often the first printed vernacular works and proved to be a critical foundation for language standardization in subsequent centuries. The Slovene language fits this pattern. The first printed book in Slovene was a Protestant catechism, translated in 1551 by Primož Trubar, a Protestant preacher. This left the printed record of the language for writers in the nineteenth and twentieth centuries, who modernized Slovene with the publication of newspapers, books, and periodicals (Biggins and Crayne 2000). Given that literate ethnic Slovenes spoke German rather than Slovene which was considered a peasant tongue for the centuries of Habsburg rule, the earlier translation effort allowed the Slovene language to survive. Similar examples can be found in Estonian, Latvian, Livonian, Lithuanian, and Finnish (Steinberg 1974, p. 122). The conscious-choice process offers an alternative path to language standardization.

Figure 2 shows the cumulative publications of vernacular dictionaries between 1750 and 2000 among the 104 ethnic groups in my data set broken down by those groups, with a record of print adoption and those without it. It indicates that no matter the route, access to the printing press gives ethnic groups a strong advantage in consolidating the vernacular.

The Significance of the Early Adoption of the Press

It is imperative to understand the timing of access to the printing press because language standardization is a slow process. It easily spans a few

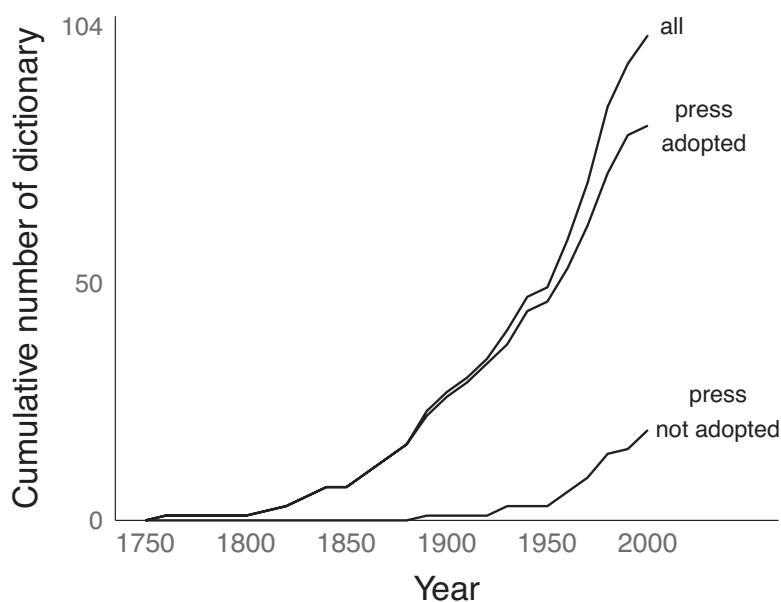


FIGURE 2
CUMULATIVE NUMBER OF VERNACULAR DICTIONARY PUBLICATION AMONG
104 EUROPEAN ETHNIC GROUPS

Source: Burke (2004), Dalby (1998), Price (1998), among others.

centuries, even under ideal circumstances in which an ethnic group had state resources and institutions at its disposal specifically to develop its vernacular. Lexicography, in the monolingual edition, entails three steps: collecting words, “tokenizing” or making a tangible representation of words, and making an entry of these words (Brown 2006b, p. 113). For each entry, lexicographers need to complete several tasks: (1) orthography (the rules about spelling); (2) guidance on pronunciation; (3) the classification of word-class (e.g., noun, verb); (4) definition making; (5) examples; (6) phraseology; (7) disputed points of usage; and (8) etymology and word histories (Brown 2006b, p. 113).

This is an ideal level of progress for language standardization. Codified tongues play a central role in commanding authority, prestige, and legitimacy when ethnic entrepreneurs seek to mobilize the community and invoke a distinct, “national” identity. Language can have these functions. Advocates for codification typically used the rationale to “purify” the vernacular, separated from the words which were characterized as “dirt” or “chaff,” as their etymology could be traced to foreign tongues (Burke 2004, pp. 144–50). Such mingling occurred because pre-standardized vernaculars were a mix of “indigenous” words which originated

from the local language and “foreign” words which were imported from other locales as a result of trade or geographical proximity, or evolved from older languages like Latin. Standardization has this distinct attribute in comparison to other dimensions of language development such as grammar-making, which primarily concerns setting up the structural rules that govern the language.

The two processes I hypothesized earlier each have a distinct route to language standardization. First, the selection process stresses state capacity to promote vernacular codification. Ethnic groups with a sovereign state started in the late sixteenth century to establish a state-funded academy devoted to studying the vernacular, following the model of the Italian academy, *Accademia della Crusca*, founded in Florence in 1584. The state-centric approach is historically a phenomenon of continental Europe. Ethnic groups that adopted this model include the French, Spanish, Portuguese, Swedish, Russian, and Danish, among others. Their centuries-long effort culminated in delivering monolingual dictionaries as the authoritative source of language use in the nineteenth and twentieth centuries (the Italian dictionary was published in 1861).

The second process focuses on actors’ conscious choice to develop the vernacular. Given the high fixed costs of codification, ethnic groups have to secure access to patrons. Such an investment grants specialists time to concentrate on their project. The codification of the English language fits this model. Unable to secure royal patronage, Samuel Johnson managed to make contracts with five booksellers, for the sum of £1,575, to cover the expenses (Brown 2006a, p. 130).¹³ Johnson’s *Dictionary of the English Language*, published in 1755, has been widely regarded as the standard-bearer. This illustration points to a language standardization process alternative to the state-centric one.

EMPIRICAL STRATEGY

I hypothesize that the variation on language standardization depends on the timing of the acquisition of the printing press for each ethnic group. Early adoption affords time to make literacy more accessible and develop a richer lexicography, by creating a specialized state agency or seeking

¹³ Johnson opposed a state-funded English academy whose model was adopted in parts of continental Europe as contrary to “the spirit of English liberty” (Landau 2001, p. 54). The English founded an academy (the Royal Society of London for the Improvement of Natural Knowledge) in 1662 and did launch a committee that addressed the grammar and orthography. But the main mission of the academy remained scientific and the committee on language met only four times and disbanded (Martin 2008, p. 225).

patronage to cover the fixed cost of the codification effort. By contrast, late adopters of the press do not have this luxury of time. If they already possess a state and resources, they might not have as strong an incentive to codify their own vernacular as early adopters. Latecomers could opt to accept the more developed language of one of their neighbors as a solution, which likely reduces the possibility to construct a distinct group of their own.

To test my argument, I constructed a new data set that contains information on various dimensions related to language standardization. It is unique in that I collect historical data on European ethnic groups as the unit of analysis. The data set is composed of 171 existing ethnic groups in Europe and that are observed for the period between 1400 and 2000 CE.¹⁴ I draw primarily on Minahan (2000) for entries on these ethnic groups and add others referred to, but not entered, in Minahan's volume. I double-check ethnic groups against *Ethnologue* compiled by M. Paul Lewis, Gary F. Simons, and Charles D. Fennig (2013). The list of 171 ethnic groups is given in the Online Appendix. The starting point is 1400 and I record data at the 50-year interval, unless otherwise noted. To enable the observation of covariates at the ethnic-group level, I locate a "homeland" city for each group and find information on political, economic, social, and geographical dimensions for these cities. I regard these observations as specific attributes for ethnic groups.¹⁵ To date, this is one of the most detailed data sets on European ethnicity and nationalism. Figure 3 displays the geographical location of the homeland cities for 171 European ethnic groups.

The outcome variable is the timing of language codification. As I defined the term "nation" by the standardization of cultural attributes for ethnic groups, this coding strategy allows me to capture the standardization process of a core attribute of ethnicity, language, and draw

¹⁴ I follow Price (1998) who relies on the geographical, geological, and historical criteria to define "Europe." Europe geographically covers land from Iceland as the westernmost border to the Ural Mountains in Russia as the easternmost border. The southeastern limit is at the Caucasus Mountains, including Georgia, Armenia, and Azerbaijan, but excluding Iran. Turkey is included. Two islands offshore Portugal, the Azores and Madeira, are included because Azores is geologically on the ridge of the European continent; while Madeira is geologically on the African continent, scholarly convention is to include it (for the cultural reason that Portuguese is the only language spoken there). Likewise, Malta is included for geological and practical reasons, although Cyprus is not, because geographically it is closer to Syria, a non-European country, than to Europe.

¹⁵ This coding strategy is both theoretical and practical. Theoretically, virtually all ethnic groups identify a city as their unique "homeland" which is shared with no other groups. Ethnic groups typically base their national movement in the homeland to claim territorial autonomy (Smith 1986). Practically, treating ethnic groups to be discrete allows me to capture a geographical impact on them. I mainly rely on Minahan (2000) for the entry of the ethnic homeland.

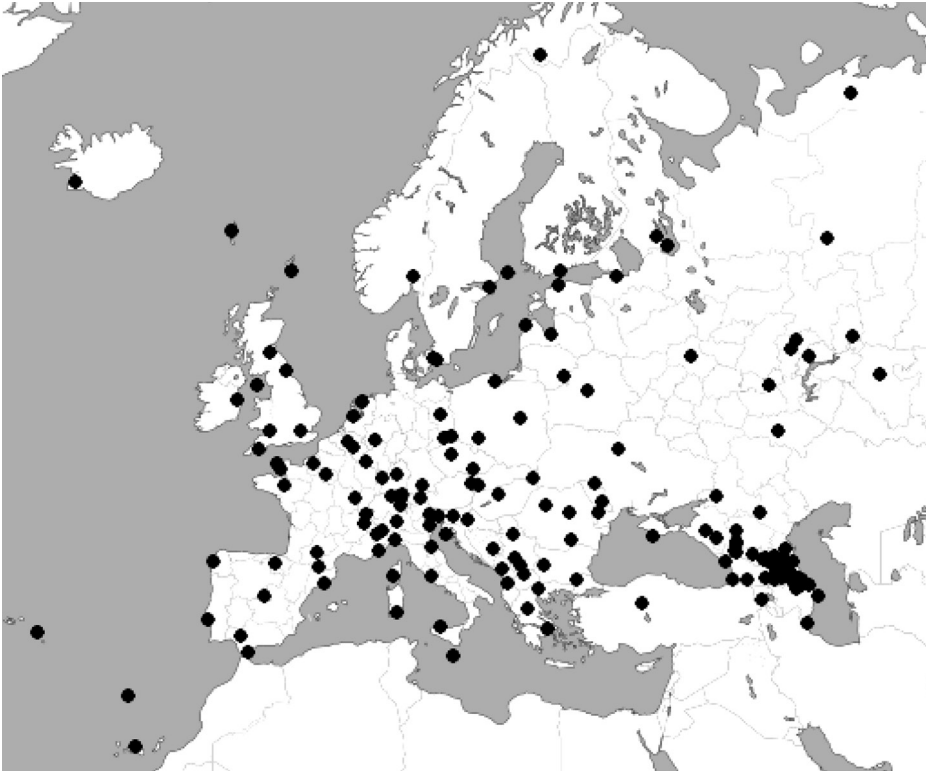


FIGURE 3
GEOGRAPHICAL DISTRIBUTION OF THE HOMELAND CITIES FOR 171 EUROPEAN
ETHNIC GROUPS

Note: The ethnic groups in the bottom left are on the islands of the Canaries, Madeira, and Azores. Those found in the north of Britain are on the islands of Shetland and the Faroes.

Source: Minahan (2000) and Lewis, Simons, and Fennig (2013).

implications for how ethnic groups consolidate their cultural practices. I operationalize the timing of language standardization by taking the first publication year of a comprehensive vernacular dictionary for each ethnic group. The qualifier “comprehensive” refers to a “modern,” monoglot dictionary that aims to cover most, if not all, words in the alphabet of the given language and one judged as setting a linguistic standard by scholars in relevant fields such as lexicography, linguistics, and cultural history. As Sidney I. Landau (2001) points out, these comprehensive dictionaries are distinct in purpose, substance, and scope from glossaries and encyclopedias, although the latter may often bear the label “dictionary” in the title.¹⁶ In a similar vein, this qualifier excludes polyglot dictionaries

¹⁶ A glossary usually contains an alphabetical list of terms used in a specific topic. An encyclopedia is a collection of entries on a given branch of knowledge.

such as the premodern translation works between a vernacular and Latin. Wherever possible, I try to find a dictionary specifically compiled for minority ethnic groups whose vernaculars today may be classified as “dialects.”¹⁷ The dictionary data rely mainly on Peter Burke (2004), Andrew Dalby (1998), and Glanville Price (1998). For some ethnic groups in the Caucasus, I use Sebastian Nordhoff, Harald Hammarström, Robert Forkel, et al. (2013). As Table 1 shows, only a quarter of the observed data ($n = 26$) achieved standardization by the twentieth century, while the rest codified their tongue within the last century, especially the latter half. The data suggests that language standardization is a modern phenomenon. For many ethnic groups that are demographically small and stateless, it is still an ongoing project. Figure 4 exhibits the geographical distribution of vernacular dictionaries between 1800–2000.

The main predictor is the timing of acquisition of the movable-type printing press. I operationalize it by recording the first date that the technology arrived in the homeland city for each ethnic group. In this process, I make a distinction between the year when the Gutenberg press was adopted in a homeland and the year when vernacular books were printed. These two sets of years are not always identical and the latter typically occurs later. In this article, I choose the former because one of the empirical goals is to assess the effect of accumulated time of technological access for a given ethnic group. The print variable comes from various sources, but primarily from Colin Clair (1976) and Lucien Febvre and Henri-Jean Martin (1976), which cover Western, Southern, and Central Europe.¹⁸ For the ethnic groups in the German-speaking area, I use Christoph Reske (2007).

I assess the relationship between the printing press and a vernacular dictionary by taking the following steps. First, I evaluate the demand-side mechanism, in which social and economic developments shape demand for the printing press. As discussed earlier, printers were essentially capitalists who were willing to go to locales likely to yield a higher return on their investments. This mechanism captures whether such pre-Gutenberg press activities determine the acquisition of the press and subsequent language standardization. Drawing from recent research in economic and political history, I control for the following set of variables. The first is the university. Universities would benefit from an ability to mass-produce books and other printed material to promote literacy. I rely on Henry C. Darby and Harold Fullard (1970) and Walter Rüegg (1992–2011, 4

¹⁷ For example, according to Dalby (1998), the Walloon language for the Walloons is classified as a “regional form” of French, and so is the Flemish language for the Flemish under Dutch.

¹⁸ I also use Burke (2004), Price (1998), and Steinberg (1974) as supplementary sources.

TABLE 1
SUMMARY STATISTICS

	<i>N</i>	Mean	Median	Standard Deviation	Minimum	Maximum
<i>Outcome Variable</i>						
Year of vernacular dictionary	104	1935	1954	51	1755	2005
<i>Explanatory Variable</i>						
Year of printing press adoption	96	1566	1526	115	1460	1930
<i>Control Variables</i>						
Year of vernacular Bible	144	1789	1837	193	1380	2012
University	67	1672	1735	208	1150	1925
Bishop	99	1250	1400	608	100	2000
Elevation (m)	171	405.3	145	581	-24	2,823
Terrain ruggedness (m)	171	137.51	59.87	205	0	1,125
Urban potential	171	8.43	2.89	14.99	0.35	99.45
War frequency (in 50-year periods)	171	8.05	5	8.75	0	37
Distance from Wittenberg (km)	171	1,669	1,429	948	89	5,292
Distance from Zürich (km)	171	1,711	1,516	1,048	0	3,350
Distance from Mainz (km)	171	1,752	1,509	1,078.7	140	5,545
Hub Roman road	171	0.12	0	0.33	0	1
Roman road	171	0.32	0	0.46	0	1
Oceanic port	171	0.38	0	0.49	0	1
Island	171	0.1	0	0	0	1
Russian Empire/Muscovy	171	0.15	0	0.36	0	1
Ottoman Empire	171	0.105	0	0.31	0	1
Habsburg Empire	171	0.08	0	0.27	0	1
Orthodox	171	0.16	0	0.37	0	1
Catholic	171	0.35	0	0.47	0	1
Protestant	171	0.17	0	0.1	0	1
Islam	171	0.29	0	0.45	0	1
Western Europe	171	0.16	0	0.36	0	1
Northern Europe	171	0.14	0	0.35	0	1
Southern Europe	171	0.16	0	0.40	0	1
Eastern Europe	171	0.40	0	0.49	0	1
Western Asia	171	0.10	0	0.30	0	1

Source: See the Empirical Strategy section for the description of each variable.

Volumes) on the history of European universities to obtain the founding date. The second control is the bishopric. Like universities, the printing press would make the proselytization effort easier and efficient, with the capacity to print pamphlets, posters, and booklets on a large scale. I use David M. Chaney (2015) for the establishment year of a diocese or archdiocese. For the university and bishopric, I record the year of foundation. The third control is urban potential. It was originally constructed by Jan

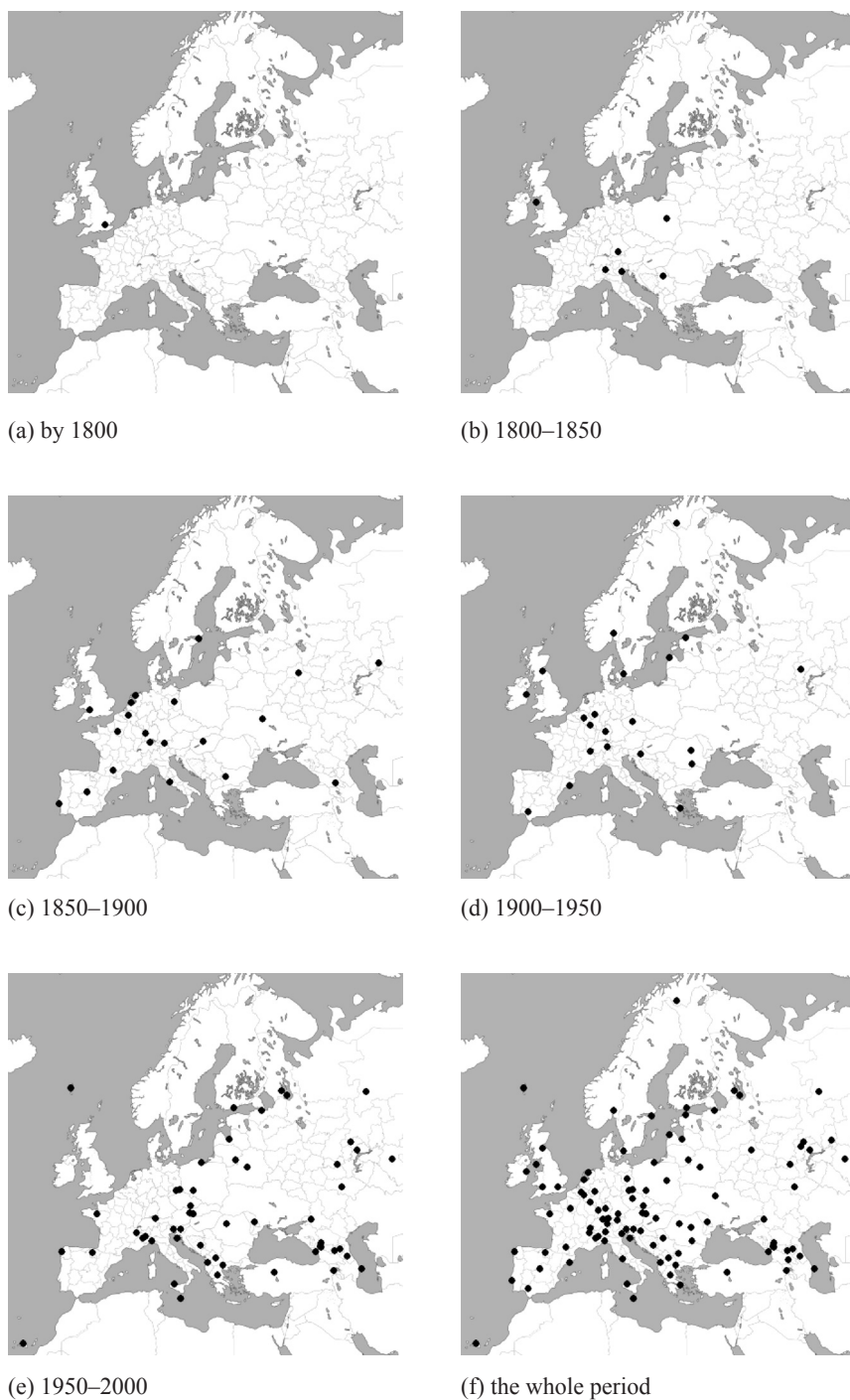


FIGURE 4
GEOGRAPHICAL DISTRIBUTION OF 104 VERNACULAR DICTIONARIES, 1800–2000.

Source: See Figure 2.

de Vries (1984) as the sum of the population in the given period divided by the geographical distance between a city and all others in my data set.¹⁹ This variable gives a sense of whether a city is surrounded by competing urbanizing towns or located in a more sparsely-populated area. Higher values indicate greater potential for urbanization. The standard source on population size in preindustrial Europe is Paul Bairoch, Jean Batou, and Pierre Chévre (1988), which covers the period 800 through 1850 for hundreds of European cities. Maarten Bosker, Eltjo Buringh, and Jan Luiten van Zanden (2013) correct some of the data in Bairoch, Batou, and Chévre, and I follow their updates in compiling my data. For the period after 1850, I use several statistical handbooks including Brian R. Mitchell (2003a, 2003b). The population data are available at the 100-year interval, so I take the average to compute values for 50-year periods.

A second set of controls concerns the supply-side mechanism. These variables can accelerate or delay language rationalization. A major covariate is war. Early-modern European history is characterized by recurring warfare, in which growing costs of fighting and preparing for war largely determined how to organize state entities most effectively (Bean 1973; Tilly 1975, 1985, 1992). A unitary language may emerge as desirable in this process. If tax collectors and subjects who are taxed communicate in a mutually intelligible tongue, this renders the administration of raising resources and manpower more effective. War can, therefore, serve as a catalyst that spurs the incentive for language rationalization. I draw on the database compiled by Peter Brecke (1999) for war-related data. It records any conflict in the world with the minimum of 32 casualties in the period between 1400 and 2000. As Brecke uses states as the unit of analysis, I take care to localize the incidence of war at the ethnic-group level to the extent possible. If, for example, a war took place in Scotland prior to union with England, I regard it as having an impact on the Scots but not others in their neighbor (the English, Welsh). However, my motive here is to capture the institutional effect of war on taxation and governance. Therefore, if no mention is made about the place of a given war, I determine that war uniformly affects the residents in the country. For instance, if a war occurred in the Habsburg Empire, I regard that war as affecting all ethnic groups within the imperial domain. I construct war frequency, which measures how often an ethnic group experiences war in any 50-year period, to capture war's impact.

Another supply-side covariate is overseas trade. If an ethnic group's homeland city is located on or near the coast, oceanic trade can promote

¹⁹ More formally, this variable is obtained by $\sum_{j=1}^N \text{population}_j / \text{distance}_{ij}$.

language standardization for greater efficiency. In European history, geography played an important role in economic activity.²⁰ Because roads for overland trade were not well-paved and thus proved unreliable, access to ports provided a critical precondition for economic growth. This access, in turn, may provide incentives for a unitary language. I use Christos Nüssli (2011) to produce an indicator taking the value of one if an ethnic group's homeland lies on an oceanic coast.²¹ While access to trade is favorable to growth, other geographical conditions may have the opposite effect. In particular, "bad terrain" may prove prohibitively costly for undertaking vernacular codification. To assess this impact, I include a series of time-invariant measures. One is a set of variables such as terrain ruggedness and land elevation above the sea level for each ethnic group. I obtain these observations from the Global Land One-kilometer Base Elevation project (GLOBE) database (GLOBE Task Team and others 1999).²² Related is a measure on island, an indicator taking the value of one if an ethnic group's homeland is on an island. All these geographical variables are intended to capture different dimensions of geography's impact.

In addition, I include indicators on the Russian, Ottoman, and Habsburg Empires based on Nüssli (2011). They take the value of one if an ethnic locale was under any of these polities at the beginning of the century. These are designed to capture the institutional impact on the acquisition of the printing press. Historians indicate that Russians and Ottomans in particular had centralized control over private, vernacular print until the eighteenth century.²³ Thus it is expected that the ethnic groups under Russian or Ottoman rule would be late in adopting the printing press and consequently in standardizing their vernacular.

Finally, I use two indicators to capture potential long-term institutional effects. One is a set of variables for the Roman Empire's influence. The Romans built roads between cities, which gave these cities an opportunity to develop institutional capacity in economic activity and political organization. I use Richard Talbert (2000), Nicholas G.L. Hammond

²⁰ See Acemoglu, Johnson, and Robinson (2005) and Bosker, Buringh, and van Zanden (2013) for the importance of geography on long-term economic growth. See Stasavage (2010) for the role of geography on the frequency of parliamentary sessions.

²¹ I exclude a similar indicator for the river, as virtually all ethnic groups founded their homeland near the river.

²² Values for terrain ruggedness in the GLOBE data set come from the Terrain Ruggedness Index, originally devised by Riley, DeGloria, and Elliot (1999).

²³ An exception includes some cities in the west of Istanbul in the Ottoman Empire, where Jewish printers were allowed to set up non-Turkish and non-Arabic presses, and some ethnic groups in the Balkans, where Ottoman rule was relatively weaker and religious legitimacy was less effective than in the Anatolian heartland (Coşgel, Miceli, and Rubin 2012).

(1981), Johan Åhlfeldt (2015), and *Pleiades* (2015) to collect information on the Roman legacy. More specifically, I create an indicator taking the value of one if a city had major or minor Roman roads.²⁴ Another institutional effect I control for is the impact of the Protestant Reformation. The data set includes two measures. One is the measure on geographical distance to Wittenberg or Zürich, two crucial cities in understanding the movement. Following Steven Pfaff and Katie E. Corcoran (2012), I construct this variable by taking the nearest distance between the given ethnic homeland and either of the Protestant centers. It is important to note that this variable captures not the impact of the confessional movement, but the magnitude of the religious reform on institutions. One of the Reformation's consequences is that cities and other localities were forced to enact institutional reforms to reduce tension after being forced to choose a side. One example is education reform for the upper-class and university education (Gorski 2003, p. 19). This implies that the Reformation could promote vernacular literacy and a demand for a linguistic standard. Proximate location to Wittenberg or Zürich indicates a greater impact of such institutional reform. The second measure on the Reformation is a fixed effect on the predominant religious preference (Catholicism, Protestantism, Orthodoxy, or Islam) for each group.²⁵ Geographical indicators for the region are also included in the data set.

ESTIMATION RESULTS

Baseline Regression

To begin, I document the baseline correlation between the printing press (the explanatory variable) and vernacular dictionaries (the outcome variable). The bivariate relationship is shown in Table 2. Of the 171 observations, approximately half (48.5 percent) acquired the printing press and achieved vernacular codification. For those ethnic groups that have the recorded date of print adoption, 86 percent standardized the vernacular by 2000 (the cutoff year for this study). By contrast, for those observations without access to the technology, 72 percent had not achieved standardization by 2000. It is important to highlight that language standardization is an ongoing project. Although 67 groups did not publish

²⁴ Following Bosker, Buringh, and van Zanden (2013), those cities with two or more major roads are coded "Roman hub" and those with one major road or one or more minor roads are coded cities with "Roman roads."

²⁵ For those ethnic groups that adopted Protestantism after the Reformation (before 1500), I code them to be associated with Catholicism.

TABLE 2
TWO-BY-TWO TABLE ON THE RELATIONSHIP BETWEEN PRINT TECHNOLOGY
AND VERNACULAR DICTIONARY (N = 171)

		Language Standardization	
		Yes	No
Printing Press	Yes	83 (48.5%)	13 (7.6%)
	No	21 (12.3%)	54 (31.6%)

Source: See Figures 1 and 2.

monolingual dictionaries and are defined as “unstandardized,” they may be able to codify their vernacular in the future.

Second, I discuss the timing of printing press adoption. Of the 83 ethnic groups that got the press and language standardization, elapsed time between the two is, on average, 363 years (the median is 396 years). There is a huge lag because 75 percent of those groups that acquired the press did so by the early seventeenth century, while 75 percent of those that standardized language did so in the twentieth century. These two pieces of descriptive evidence provide preliminary support for my hypotheses that the adoption of the printing press is positively linked to language standardization and that, given this lag, the early adoption of presses is crucial to it.

To test my argument more systematically, this article employs the following estimation strategies. First, I use the Cox proportional hazard model to examine the effect of time-varying covariates on vernacular codification. The key advantages of the Cox model include the assumption that the baseline hazard rate does not follow a particular distribution. Instead, the duration times are parameterized in terms of a given set of covariates (Box-Steffensmeier and Jones 2004, p. 49). The positive coefficients are interpreted as increasing hazards for an event of interest as a function of the covariates, namely increasing the chances of language standardization for ethnic groups. By contrast, the negative coefficients are interpreted as decreasing hazards, meaning that the chances of language rationalization decrease. More specifically, I estimate the following reduced-form model:

$$h_i(t) = h_0(t) \exp(\beta_1 \text{Press}_i + \gamma_1 X_i + \delta_i). \tag{1}$$

Here $h_i(t)$ is the codification of vernacular dictionaries for ethnic group i at time t . This event of interest is parameterized by whether ethnic groups acquired the printing press (β 's), a vector of covariates γX_p , and a set of fixed effects δ for empires, geography, and religion. $h_0(t)$ is the baseline

hazard but dropped in estimation. In addition to the Cox model, I use logistic regression and estimate the same specifications as robustness checks.²⁶

Table 3 documents regression estimates for the Cox and logistic models, each in four sets of specifications: (1) the bivariate model; (2) the model with demand-side covariates; (3) the model with supply-side covariates (the demand-side ones included); and (4) the fully-specified model that includes all the controls and fixed effects. The bivariate model suggests that if ethnic groups acquire a printing press in the homeland, the chances that they standardize the vernacular is 8.66 times greater than those without access to the technology (the value of the hazard ratio obtained by $\exp(2.159)$ in Model 1). The result holds when the demand-side covariates are introduced. The magnitude of the hazard ratio gets attenuated to 4.95 in Model 3, with the founding of a university significantly positively correlated with language standardization. Other specifications exhibit a similar pattern. The supply-side scenario indicates that rough terrain and war are expected to have a significantly negative impact on language standardization, although the substantive effect of the press remains approximately the same at 4.29 (Model 5). Finally, the magnitude of the press is invariant to a host of fixed effects on imperial rule, religion, and geographical region. In all specifications, the adoption of print technology is significantly positively correlated with vernacular codification and the result is robust to the inclusion of a host of covariates and fixed effects.

A second hypothesis is that the early acquisition of the printing press is crucial for the early codification of the vernacular. To test this proposition, I create counterfactual scenarios where ethnic groups acquired the press at the different timings and estimate the cumulative hazard for language standardization. I set three scenarios in the 150-year interval, 1500, 1650, and 1800 CE, and estimate the cumulative hazard from the fully-specified Cox model. Cumulative hazard rates denote the total amount of risk that has been accumulated over time. This exercise shows that the advantage of early adoption become discernible over time. The cumulative hazard rate reaches 0.2 for early adopters (presses by 1500), meaning that the chances of language standardization are estimated to add up to 0.2 if ethnic groups acquired print by 1500. The estimated cumulative hazard is 0.16 for “mid”-adopters (presses by 1650) and 0.14 for late adopters (presses by 1800). Though the differences in magnitude are not large, the simulation offers two insights. First, early technology adoption can

²⁶ The logistic models include the time trend and its squared function as suggested by Carter and Signorino (2010). The complete estimates discussed in this article all are reported in the Online Appendix.

TABLE 3
REGRESSION OUTPUTS OF THE PRINTING PRESS' IMPACT ON LANGUAGE STANDARDIZATION

Model	Bivariate		Demand Side		Supply Side		Fully Specified	
	Cox (1)	Logit (2)	Cox (3)	Logit (4)	Cox (5)	Logit (6)	Cox (7)	Logit (8)
Printing press	2.159*** (0.269)	2.933*** (0.342)	1.600*** (0.356)	2.178*** (0.456)	1.455*** (0.354)	1.932*** (0.503)	1.450*** (0.339)	2.083*** (0.556)
University			0.0004*** (0.0001)	0.001*** (0.0002)	0.0004*** (0.0001)	0.001*** (0.0003)	0.0004*** (0.0001)	0.001*** (0.0002)
Bishop			-0.000002 (0.0001)	-0.0001 (0.0002)	0.00003 (0.0001)	0.00004 (0.0002)	-0.00002 (0.0001)	-0.00002 (0.0003)
Log urban potential			0.114 (0.226)	0.098 (0.365)	-0.002 (0.607)	-0.307 (0.793)	0.186 (0.636)	-0.276 (0.821)
Log distance to Wittenberg/Zürich					0.0001 (0.0003)	-0.0001 (0.001)	0.0004 (0.0003)	0.0002 (0.001)
War frequency					-0.031** (0.014)	-0.043*** (0.016)	-0.030** (0.013)	-0.048*** (0.018)
Elevation					-0.0001 (0.0004)	-0.00002 (0.001)	0.0001 (0.0004)	0.0005 (0.001)
Terrain ruggedness					-0.003** (0.001)	-0.004** (0.002)	-0.002** (0.001)	-0.004** (0.002)
Hub Roman road					0.156 (0.296)	0.494 (0.634)	0.129 (0.337)	0.431 (0.7)
Roman road					-0.009 (0.247)	-0.023 (0.467)	0.081 (0.304)	0.151 (0.603)

TABLE 3 (CONTINUED)
REGRESSION OUTPUTS OF THE PRINTING PRESS' IMPACT ON LANGUAGE STANDARDIZATION

Model	Bivariate		Demand Side		Supply Side		Fully Specified	
	Cox (1)	Logit (2)	Cox (3)	Logit (4)	Cox (5)	Logit (6)	Cox (7)	Logit (8)
Oceanic port					-0.415** (0.212)	-0.712* (0.411)	-0.477* (0.26)	-0.721 (0.539)
Island					-0.057 (0.475)	0.185 (0.85)	-0.038 (0.52)	0.2 (0.965)
Empire fixed effects	No	No	No	No	No	No	Yes	Yes
Religion fixed effects	No	No	No	No	No	No	Yes	Yes
Geographic region fixed effects	No	No	No	No	No	No	Yes	Yes
Time dummies		Yes		Yes		Yes		Yes
Observations	2,052	2,052	2,052	2,052	2,052	2,052	2,052	2,052

* = Significant at the 10 percent level.

** = Significant at the 5 percent level.

*** = Significant at the 1 percent level.

Notes: Robust standard errors clustered by ethnic groups for all models. Intercept in the logistic models is not reported. *Western Europe* and *Catholic* are used as the reference category, respectively, for geographic region fixed effects and for religion fixed effects so they are omitted. Full results are reported in the Online Appendix.

Source: See the Empirical Strategy section.

trigger knowledge accumulation. Vernacular printing leaves a written record of cultural practices for ethnic groups. The stock of knowledge in a unique language grows over time through access to concrete, verifiable information in printed material (Mokyr 2002, 2005). Printing technology can facilitate cultural innovation by lowering the costs of writing novels, mythologies, and history, which in turn consolidates language. Second, this simulation suggests that increasing efficiency in technology may not permit quick catch-up in cultural rationalization for late adopters. Although the costs associated with printing were expected to have fallen over this period, the challenge of assembling knowledge on culture does not appear to be quickly overcome. Estimates from these scenarios indicate that the timing of acquiring the enabling technology matters for understanding language standardization.

Robustness Checks: The Role of the State

Regression estimates so far supported my hypotheses that the adoption of printing technology is positively correlated with language standardization and that the timing of adoption is crucial. However, there are two endogeneity concerns to my causal claims. One is the impact of independent statehood. This channel represents the selection process, in which ethnic groups with access to an independent state were disproportionately endowed with resources and capabilities that accelerate language standardization. This process could have begun before 1400 CE, the year when my analysis time starts, and print adoption is endogenous to this process. For instance, although the literature suggests that the spread of the press depended primarily on skilled workers (Cipolla 1972), an early start on building state capacity in taxation or rule enforcement can determine the demand for the printing press, because the technology can make governance more efficient. It can thus be imagined that the acquisition of print technology is a function of having an independent state. To address this concern, I employ the following empirical strategy. I begin by constructing an alternative data set with the state as the unit of analysis ($n = 47$). Within this sample, I recode the printing variable by identifying its first adoption year for any city in each state. I then reestimate the impact of printing technology on vernacular codification. For states with multiple official languages, I use the first publication date of a vernacular dictionary for each tongue and use it as an alternative outcome (my list has up to four official languages). The Online Appendix offers the list of the new sample with key variables. Table 4 presents the Cox regression estimates for the state sample.

TABLE 4
COX PROPORTIONAL HAZARDS MODEL OUTPUTS OF THE PRINTING PRESS'
IMPACT ON LANGUAGE STANDARDIZATION

Model	State				Non-state
	List 1 (1)	List 2 (2)	List 3 (3)	List 4 (4)	(5)
Printing press	14.711*** (0.98)	14.984*** (0.92)	14.725*** (1.1)	14.814*** (0.97)	4.915** (2.293)
All controls	Yes	Yes	Yes	Yes	Yes
Observations	564	564	564	564	1,488

* = Significant at the 10 percent level.

** = Significant at the 5 percent level.

*** = Significant at the 1 percent level.

Notes: Robust standard errors clustered by ethnic groups for all models. *Western Europe* and *Catholic* are used as the reference category, respectively, for geographic region fixed effects and for religion fixed effects so they are omitted. Full results are reported in the Online Appendix.

Source: See Figure 1 for the sources of the printing press.

All models include the covariates from the fully-specified model. The impact of the first printing press on language standardization remains significantly positive. The magnitude of coefficients is stable across alternative lists of official languages and much greater than that in the ethnic-group sample. This is unsurprising, because the vast majority of 47 states have the recorded date of print adoption and have standardized the vernacular. At the same time, if unobserved state-related forces determine language standardization, their impact on those ethnic groups without access to the state is expected to be statistically indistinguishable from zero. Column 5 of Table 4 reports estimates of Cox regression for these groups ($n = 124$). Although the substantive effect of print technology is much smaller than in the state-only sample, the results remain similar to Table 3. This simulation suggests that while the selection process in which the adoption of the press is a function of having a state may be at work, the unobserved state capacity is unlikely to determine the chances of language standardization.

Robustness Checks: Human Capital's Impact

The second endogeneity concern is the influence of human capital on the development of the vernacular. While the urbanization potential is controlled for as a proxy for economic growth, a more specific channel through human capital, or individual-level literacy, may account for language standardization. In the literature on economic history and on

modern-era growth, strong cognitive skills can enhance economic well-being.²⁷ Increase in literacy is hypothesized to have a spillover effect on an investment in the vernacular for greater efficiency in communication, writing, and business transactions within ethnic-group members. In the context of early-modern Europe, a major driver for human capital is the Protestant Reformation (Becker and Woessmann 2009, 2011; Boppart, Falkinger, and Grossmann 2014; Cantoni, Dittmar, and Yuchtman 2016).²⁸ The religious movement prodded lay followers to read the Bible in the vernacular. Literate reformers also issued church ordinances to reinforce lay literacy (Dittmar and Meisenzahl 2016). Recent research demonstrates that an (early) access to the printing press played a critical role, because campaigners of the religious reform would take advantage of the technology's capacity to mass-produce, disseminate information, and canvass support through the vernacular Bible, ordinances, and broadsheets (Rubin 2014). If the human capital hypothesis is correct, ethnic groups that observe Protestantism as the primary religion are expected to exhibit high literacy, which in turn is positively correlated with language standardization. The adoption of the printing press may be endogenous to this process.

This article has already addressed the impact of the Protestant Reformation on print and vernacular codification by including the religion fixed effects and taking the shorter distance from an ethnic group homeland to either Wittenberg or Zürich. Yet these measures may be too broad to capture the human capital channel precisely.

To account for this channel, I construct the vernacular Bible indicator. It is an appropriate proxy for a source of human capital in early-modern Europe, because it captures the Protestant advocacy of lay literacy built on the access to printing technology. This variable takes the first publication date (year) of the vernacular translation of the Bible drawn from *Ethnologue* which is compiled by Lewis, Simons, and Fennig (2013). The data is also supplemented by Price (1998). As with the dictionary publication, the year 2000 is used as the cutoff. My strategy is, first, to estimate the impact of the vernacular Bible in the fully-specified Cox and logistic models *minus* the print variable (i.e., Columns 7 and 8 in Table 3, respectively). Second, I reintroduce the print variable with the Bible variable included. If the human capital channel determines the chances of

²⁷ For general works on the role of human capital in growth, see, for instance, Easterlin (1981), Goldin (2001), Lindert (2003), Glaeser and Saiz (2004), Algan and Cahuc (2014), Hanushek and Woessmann (2008).

²⁸ Becker, Pfaff, and Rubin (2016) overviews major theories and empirical works. For a negative association between the religious reform and growth, see Cantoni (2015).

TABLE 5
REGRESSION OUTPUTS OF THE PRINTING PRESS' IMPACT
ON LANGUAGE STANDARDIZATION

Model	Cox (1)	Logit (2)	Cox (3)	Logit (4)
Printing press			1.394*** (0.342)	2.023*** (0.558)
Vernacular Bible	0.531** (0.251)	0.745** (0.356)	0.461* (0.257)	0.676* (0.365)
All controls	Yes	Yes	Yes	Yes
Observations	2,052	2,052	2,052	2,052

* = Significant at the 10 percent level.

** = Significant at the 5 percent level.

*** = Significant at the 1 percent level.

Notes: Robust standard errors clustered by ethnic groups for all models. Intercept in the logistic models is not reported. *Western Europe* and *Catholic* are used as the reference category, respectively, for geographic region fixed effects and for religion fixed effects so they are omitted. Full results are reported in the Online Appendix.

Source: See Figure 1 for the sources of the printing press and see Lewis, Simons, and Fennig (2013) and Price (1998) for the vernacular Bible.

language standardization, the vernacular Bible should be positively associated with language standardization. Moreover, the substantive impact of the printing press is expected to be statistically indistinguishable from zero, while the Bible's effect should be retained.

Table 5 documents the role of human capital in language standardization. As before, all models are fully-specified ones including the fixed effects. Table 5 shows that the vernacular Bible is significantly positively correlated with vernacular codification in all models. In Model 1, ethnic groups with the Bible have 1.7 times greater chances of language standardization than those without it. When the press is included in Model 3, the magnitude gets attenuated (with the chances of the event now 1.59 times) but stays positive. Yet printing technology exhibits the much greater magnitude in the same model, in which the press increases the chances of language standardization by 4 times. Even when compared to fully-specified Model 7 of Table 3, the hazard ratio for print drops only slightly by 0.26. Although the human capital channel may account for why ethnic groups standardize the vernacular, the printing press hypothesis remains robust and exhibits a greater impact.

To disentangle the regression results from Table 5, it is useful to revisit the sequence of historical events. The invention of the movable type *preceded* not just the Protestant Reformation but also Bible translation for all cases in my observations. Many European ethnic groups enjoyed

a vernacular print earlier than the translation of the Bible promoted by Protestants. Model 3 of Table 5 shows that the changes in the hazard ratio of the press is 2.53 points greater than those of the Bible, pointing to the importance of *earlier* adoption of the press for European ethnic groups. The chronological order matters when considering the differences in the magnitude of these different channels.

Distance to Mainz as an Instrumental Variable

The regression analysis thus far documents the positive association between the printing press and language standardization, which is robust to the inclusion of a host of covariates under various scenarios. However, as previously mentioned, the distribution of the printing press is not random. Unobserved (pre-press) characteristics may drive the technology's spread or jointly determine press adoption and language standardization. To account for this endogeneity concern, I follow Jeremiah E. Dittmar (2011) and Jared Rubin (2014) in exploiting the exogenous variation of the geographical distance to Mainz, Germany, as an instrumental variable for the homelands of European ethnic groups.

Distance to Mainz is an ideal instrument for my argument. Such an instrument should be correlated with the printing press, but not with vernacular dictionaries. At the same time, it should affect the outcome variable only through the proposed causal mechanism. The distance-to-Mainz variable satisfies these criteria. To begin, it has been shown that the Gutenberg press diffused through Europe in roughly a concentric-circle fashion. This describes not only the spread of the press, but also the patterns of human interactions in early-modern Europe more generally. Dittmar (2011) points out that Gutenberg and his collaborators jealously shielded the proprietary knowledge of the technology. Only nearly a century after the invention was the earliest known manual on the metal type published (Dittmar 2011, p. 1140). Geographical proximity thus offers greater chances of accessing information. More broadly, premodern times were characterized by the "small world," in which the distribution of technology or information occurred in the concentric circle manner. Using a mathematical model, Seth A. Marvel, Travis Martin, Charles R. Doering, et al. (2013) show that the bubonic plague, which triggered the Black Death and decimated approximately the two-thirds of the European population in the mid-fourteenth century, spread in this manner. Hence, the closer an ethnic group's homeland is to Mainz, the more likely it is to adopt the printing press. Figure 5, albeit not perfect, broadly supports this statement.

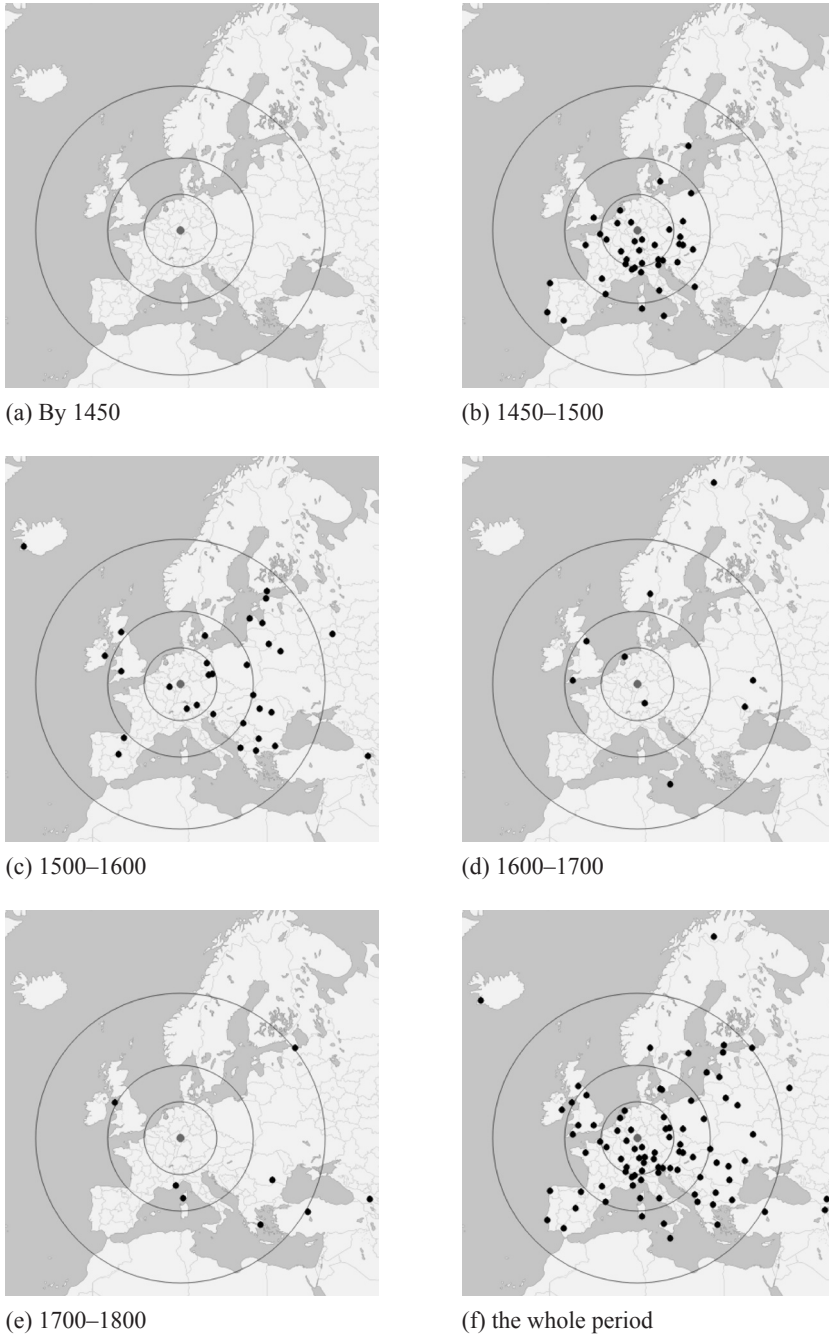


FIGURE 5
GEOGRAPHICAL DISTRIBUTION OF NET ADDITIONAL PRINTING PRESSES BY
PERIOD AND ETHNIC GROUPS, 1450–1800

Note: The gray dot denotes Mainz (the center in each map) and black dots denote the homeland city that adopted the press in the given period. The circles indicate, from small to big, the geodesic distance from Mainz at 500km, 1,000km, and 2,000km, respectively.

Source: See Figure 1.

The second rationale for using this instrument is the absence of a theoretical connection between Mainz and language standardization. No known ethnic groups have identified Mainz as their homeland city or chosen it by its geographical proximity to Mainz; nor did the city play a role in the diffusion of the vernacular use. Similarly, Rubin (2014) demonstrates that Mainz was not a political, economic, or religious center before the printing press. That Mainz was an ordinary town also implies that it is unlikely to be connected to war or economic growth. The invention may have fueled the pace of growth, but the location of Mainz is unlikely to predict it. Recent empirical works also exploit the city’s exogeneity in their estimation methods for outcomes such as the Protestant Reformation and economic growth in early-modern Europe (Dittmar 2011; Rubin 2014).

To estimate the effect of the distance-to-Mainz variable as the instrument, I estimate the IV probit model with the following system of equations:

$$\text{Press}_i = \alpha_2 + \beta_2 \text{Distance to Mainz}_i + \gamma_2 X_i + \delta_i + \varepsilon_i, \quad (2)$$

$$\begin{aligned} \text{Pr}(\text{ethnic group } i \text{ having dictionary in Year } t = 1 \mid X_i, \widehat{\text{Press}}_i) \quad (3) \\ = \Phi(\alpha_3 + \beta_3 \widehat{\text{Press}}_i + \gamma_3 X_i + \delta_i + \varepsilon_i). \end{aligned}$$

I use ordinary least square (OLS) to estimate the first stage where the outcome variable is the adoption of the printing press. In the second stage, I use a probit regression, where Φ denotes the normal cdf, by including the predicted values of the first stage model for the press. The model includes all controls from the fully specified model in Table 3, but for the first stage I recalculated the urban potential and war frequency variables to confine their effects to the fifteenth century. I do this to see if Mainz and its distance to the ethnic homelands were correlated with these variables *prior* to the Gutenberg invention. In addition, I omit the distance to Wittenberg or Zürich and replaced the Protestant with Catholic fixed effects for the first stage to follow the chronological sequence of events.²⁹ This calibration also allows me to assess whether Mainz was an economic center or underwent conflict before or around the time when the printing press was invented. In the second stage, I use the predicted values of the first-stage regression to estimate language standardization with the same set of covariates, including the distance to Wittenberg or Zürich and the Protestant dummy reintroduced. To see the impact of the printing press

²⁹ See Febvre and Martin (1976), Bernstein (2013), and Eisenstein (1979) for theoretical propositions and Rubin (2014) for empirical examination.

TABLE 6
IV PROBIT REGRESSION OUTPUTS OF THE PRINTING PRESS' IMPACT
ON LANGUAGE STANDARDIZATION

Dependent Variable	First Stage	Second Stage			
	Printing Press (1)	Language Standardization			
		By 1850 (2)	By 1900 (3)	By 1950 (4)	By 2000 (5)
Log distance to Mainz	-0.159*** (0.060)				
Printing press		11.822 (37.038)	4.639 (3.051)	4.741** (2.284)	4.018** (1.828)
All controls	Yes	Yes	Yes	Yes	Yes
<i>N</i> vernacular dictionary		7	27	48	104
Observations	2,052	1,539	1,710	1,881	2,052
<i>p</i> -value, Wald exogeneity test		0.75	0.12	0.038*	0.028*
<i>F</i> -statistic on weak instrument	24.99				

* = Significant at the 10 percent level.

** = Significant at the 5 percent level.

*** = Significant at the 1 percent level.

Notes: Robust standard errors clustered by ethnic groups. IV probit estimation: first stage is OLS, second stage is probit, regressed on predicted values from the first stage. In the first stage, *urban potential* and *war frequency* are for the fourteenth centuries. *Western Europe* is used as the reference category for region fixed effects and thus omitted. In the second stage, *Catholic* is used as the reference category for religion fixed effects and thus omitted. Full results are reported in the Online Appendix.

Source: See the Empirical Strategy section.

on early standardization, I subset the data set in accordance with vernacular codification (1) by 1850, (2) by 1900, (3) by 1950, and (4) by 2000.

Table 6 documents the results from IV probit regression. It indicates that the impact of the printing press on language standardization is positive and largely significant, when instrumented for the distance to Mainz. In the first stage, the distance-to-Mainz variable is negative and significant. The sign of the coefficient means that the closer ethnic groups' homeland is located to Mainz, the more likely they are to acquire the print technology. This result is consistent with the empirical literature that uses this instrument. The *F*-statistic on the weak instrument is approximately 25, which is above the conventional threshold of 10. In the second stage, the magnitude of the printing press is consistently in the expected direction. For the period 1850 and 1900, the printing press is positive, but not significant. This is largely because of the lack of variation in the outcome variable: by 1850 there were only seven vernacular dictionaries and by 1900, 27. But standard errors shrink as data in the outcome variable for each

period get richer. Seventy-five percent of vernacular codification took place during the twentieth century. The coefficients of the print variable after 1900 become significant in the second-stage regression, reflecting the underlying data. Notably, the magnitude remains quite large: access to printing technology increases the probability of producing vernacular dictionaries by 400–470 percentage points in the twentieth century. Despite the lack of data for early periods, the instrumental-variables approach provides additional support for my hypothesis that the printing press predicts language standardization for European ethnic groups.

CONCLUSION

I have systematically investigated the association between the printing press and the standardization of the vernacular for European ethnic groups from a long-term historical perspective. I argued that the Gutenberg press substantially reduced the cost of access to information, thus enabling vernaculars to be more popularly used and eventually win over Latin as the primary vehicle of written communication for political, economic, and social transactions. Moreover, I hypothesized that since language codification takes a long time in the order of centuries, the early acquisition of the press should give ethnic groups a head-start to develop their tongue. Using a new data set I constructed, this article has demonstrated that the time between the press and vernacular dictionaries takes 360 years on average. Statistical analysis confirms my hypotheses that print technology is positively and significantly associated with language standardization. It also supports my broader arguments that variation in cultural consolidation for ethnic groups is not solely attributed to territorial sovereignty and that historical events have a persistent impact on contemporary outcomes.

To what extent does my argument carry outside Europe? Although such an analysis is beyond the scope of this study, it seems that there are a few Europe-specific attributes relevant to language standardization. African experience, for instance, provides a useful comparative perspective. One similarity is that consistent with the European experience, recent empirical research documents that (early) access to technologies that enable human-capital development is positively associated with vernacular codification and other outcomes such as democracy in the long-run (Cagé and Rueda 2016; Woodberry 2012). However, unlike in Europe, access to printing technology in sub-Saharan Africa beginning in the nineteenth century was largely limited by proximity to the Protestant missionaries; in addition, there was no indigenous capacity to produce

the movable-type press (Cagé and Rueda 2016, pp. 73, 74). A reliance on imports suggests that the cost of access to information would remain high, leaving less room for literacy in general and the development of vernacular culture more specifically. European colonialism likely reinforced this trend. By comparison, early-modern Europe had an environment in which printers moved across Europe to spread the technology and there were multiple routes to gain access to it. This article has sought to demonstrate that such conditions were critical to standardization for many minority tongues in Europe that have survived despite institutional centralization in modern times.

REFERENCES

- Acemoglu, Daron, Simon Johnson, and James A. Robinson. "The Rise of Europe: Atlantic Trade, Institutional Change, and Economic Growth." *American Economic Review* 95, no. 2 (2005): 546–79.
- Ahlerup, Pelle, and Ola Olsson. "The Roots of Ethnic Diversity." *Journal of Economic Growth* 17, no. 2 (2012): 71–102.
- Åhlfeldt, Johan. "Digital Atlas of the Roman Empire," 2015. Access at <http://dare.ht.lu.se/>.
- Albaugh, Ericka A. *State-Building and Multilingual Education in Africa*. New York: Cambridge University Press, 2015.
- Alesina, Alberto, Arnaud Devleeschauwer, William Easterly, et al. "Fractionalization." *Journal of Economic Growth* 8, no. 2 (2003): 155–94.
- Alesina, Alberto, and Eliana La Ferrara. "Ethnic Diversity and Economic Performance." *Journal of Economic Literature* 43, no. 3 (2005): 762–800.
- Algan, Yann, and Pierre Cahuc. "Trust, Growth, and Well-Being: New Evidence and Policy Implications." In *Handbook of Economic Growth*, Vol. 2A, edited by Philippe Aghion and Steven N. Durlauf, 49–120. New York: Elsevier, 2014.
- Anderson, Benedict. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Rev. ed. London: Verso, 2006.
- Ashraf, Quamrul, and Oded Galor. "The 'Out of Africa' Hypothesis, Human Genetic Diversity, and Comparative Economic Development." *American Economic Review* 103, no. 1 (2013): 1–46.
- Bairoch, Paul, Jean Batou, and Pierre Chévre. *La Population des Villes Européennes de 800 à 1850: Banque de Données et Analyse Sommaire des Résultats (The Population of European Cities from 800 to 1850: Data Bank and Short Summary of Results)*. Genève: Librairie Droz, 1988.
- Bean, Richard. "War and the Birth of the Nation State." *Journal of Economic History* 33, no. 1 (1973): 203–21.
- Becker, Sascha O., and Ludger Woessmann. "Was Weber Wrong? A Human Capital Theory of Protestant Economic History." *Quarterly Journal of Economics* 124, no. 2 (2009): 531–96.
- . "The Effects of the Protestant Reformation on Human Capital." In *The Oxford Handbook of the Economics of Religion*, edited by Rachel M. McCleary, 93–110. New York: Oxford University Press, 2011.

- Becker, Sascha O., Steven Pfaff, and Jared Rubin. "Causes of Consequences of the Protestant Reformation." *Explorations in Economic History* 62 (2016): 1–25.
- Bell, David A. *The Cult of the Nation in France: Inventing Nationalism, 1680–1800*. Cambridge, MA: Harvard University Press, 2001.
- Bernstein, William J. *Masters of the Word: How Media Shaped History from the Alphabet to the Internet*. New York: Grove Press, 2013.
- Biggins, Michael, and Janet Crayne, eds. *Publishing in Yugoslavia's Successor States*. New York: Haworth Information Press, 2000.
- Boppart, Timo, Josef Falkinger, and Volker Grossmann. "Protestantism and Education: Reading (the Bible) and Other Skills." *Economic Inquiry* 52, no. 2 (2014): 874–95.
- Bosker, Maarten, Eltjo Buringh, and Jan Luiten van Zanden. "From Baghdad to London: Unraveling Urban Development in Europe, the Middle East, and North Africa, 800–1800." *Review of Economics and Statistics* 95, no. 4 (2013): 1418–37.
- Box-Steffensmeier, Janet M., and Bradford S. Jones. *Event History Modeling: A Guide for Social Scientists*. New York: Cambridge University Press, 2004.
- Brecke, Peter. "Violent Conflicts 1400 A.D. to the Present in Different Regions of the World." Paper presented at the 1999 Meeting of the Peace Science Society, 1999.
- Breuilly, John. *Nationalism and the State*, 2nd ed. Chicago: University of Chicago Press, 1993.
- Brown, Keith, ed. *Encyclopedia of Language and Linguistics*, Vol 6. 2nd ed. London: Elsevier, 2006a.
- , ed. *Encyclopedia of Language and Linguistics*, Vol. 7. 2nd ed. London: Elsevier, 2006b.
- Brubaker, Rogers. *Citizenship and Nationhood in France and Germany*. Cambridge, MA: Harvard University Press, 1992.
- . *Ethnicity without Groups*. Cambridge, MA: Harvard University Press, 2004.
- Buringh, Eltjo, and Jan Luiten van Zanden. "Charting the 'Rise of the West': Manuscripts and Printed Books in Europe, A Long-Term Perspective from Sixth through Eighteenth Centuries." *Journal of Economic History* 69, no. 2 (2009): 409–45.
- Burke, Peter. *Languages and Communities in Early Modern Europe*. Cambridge: Cambridge University Press, 2004.
- Cagé, Julia, and Valeria Rueda. "The Long-Term Effects of the Printing Press in sub-Saharan Africa." *American Economic Journal: Applied Economics* 8, no. 3 (2016): 69–99.
- Cantoni, Davide. "The Economic Effects of the Protestant Reformation: Testing the Weber Hypothesis in the German Lands." *Journal of European Economic Association* 13, no. 4 (2015): 561–68.
- Cantoni, Davide, Jeremiah Dittmar, and Noam Yuchtman. "Reformation and Reallocation: Religious and Secular Economic Activity in Early Modern Germany." Mimeo, 2016.
- Cardinal, Linda, and Selma K. Sonntag, eds. *State Traditions and Language Regimes*. Montréal: McGill-Queen's University Press, 2015.
- Carter, David B., and Curtis S. Signorino. "Back to the Future: Modeling Time Dependence in Binary Data." *Political Analysis* 18, no. 3 (2010): 271–92.
- Chandra, Kanchan. "What Is Ethnic Identity and Does It Matter?" *Annual Review of Political Science* 9 (2006): 397–414.

- Chaney, David M. "The Hierarchy of the Catholic Church: Current and Historical Information about Its Bishops and Dioceses," 2015. Accessed at www.catholic-hierarchy.org.
- Cipolla, Carlo M. "The Diffusion of Innovations in Early Modern Europe." *Comparative Studies in Society and History* 14, no. 1 (1972): 46–52.
- Clair, Colin. *A History of European Printing*. New York: Academic Press, 1976.
- Coşgel, Metin M., Thomas J. Miceli, and Jared Rubin. "The Political Economy of Mass Printing: Legitimacy, Revolt, and Technological Change in the Ottoman Empire." *Journal of Comparative Economics* 40, no. 3 (2012): 357–71.
- Comin, Diego, William Easterly, and Erick Gong. "Was the Wealth of Nations Determined in 1000 BC?" *American Economic Journal: Macroeconomics* 2, no. 3 (2010): 65–97.
- Dalby, Andrew. *A Guide to World Language Dictionaries*. London: Library Association Publishing, 1998.
- Darby, Henry C., and Harold Fullard, eds. *The New Cambridge Modern History*. Vol. 14: *Atlas*. Cambridge: Cambridge University Press, 1970.
- de Vries, Jan. *European Urbanization, 1500–1800*. Cambridge, MA: Harvard University Press, 1984.
- Desmet, Klaus, Ignacio Ortuno-Ortín, and Romain Wacziarg. "The Political Economy of Linguistic Cleavages." *Journal of Development Economics* 97, no. 2 (2012): 322–38.
- Dincecco, Mark. "The Rise of Effective States in Europe." *Journal of Economic History* 75, no. 3 (2015): 901–18.
- Dittmar, Jeremiah E. "Information Technology and Economic Change: The Impact of Printing Press." *Quarterly Journal of Economics* 126, no. 3 (2011): 1133–75.
- Dittmar, Jeremiah E., and Ralf R. Meisenzahl. "State Capacity and Public Goods: Institutional Change, Human Capital, and Growth in Early Modern Germany." Mimeo, 2016.
- Easterlin, Richard A. "Why Isn't the Whole World Developed?" *Journal of Economic History* 41, no. 1 (1981): 1–19.
- Easterly, William, and Ross Levine. "Africa's Growth Tragedy: Policies and Ethnic Divisions." *Quarterly Journal of Economics* 112, no. 4 (1997): 1203–50.
- Eisenstein, Elizabeth L. *The Printing Press as an Agent of Change: Communications and Cultural Transformations in Early-modern Europe*. Cambridge: Cambridge University Press, 1979.
- Ertman, Thomas. *Birth of the Leviathan: Building States and Regimes in Medieval and Early Modern Europe*. New York: Cambridge University Press, 1997.
- Fearon, James D. "Why Ethnic Politics and 'Pork' Tend to Go Together." Mimeo, 1999.
- Febvre, Lucien, and Henri-Jean Martin. *The Coming of the Book: The Impact of Printing, 1450–1800*, 3rd ed. London: Verso, 1976.
- Gellner, Ernest. *Nations and Nationalism*, 2nd ed. Malden, MA: Blackwell, 2006.
- Glaeser, Edward L., and Albert Saiz. "The Rise of the Skilled City." *Brookings-Wharton Papers on Urban Affairs*, 2004.
- GLOBE Task Team and others. "The Global Land One-kilometer Base Elevation (GLOBE) Digital Elevation Model," 1999. Accessed at <http://www.ngdc.noaa.gov/mgg/topo/globe.html>

- Goldin, Claudia. "The Human-Capital Century and American Leadership: Virtues of the Past." *Journal of Economic History* 61, no. 2 (2001): 263–92.
- Gorski, Philip S. *The Disciplinary Revolution: Calvinism and the Rise of the State in Early Modern Europe*. Chicago: University of Chicago Press, 2003.
- Graff, Harvey J. *The Legacies of Literacy: Continuities and Contradictions in Western Culture and Society*. Bloomington: Indiana University Press, 1987.
- Guiso, Luigi, Paola Sapienza, and Luigi Zingales. "Does Culture Affect Economic Outcomes?" *Journal of Economic Perspectives* 20, no. 2 (2006): 23–48.
- Hammond, Nicholas G.L., ed. *Atlas of the Greek and Roman World in Antiquity*. Park Ridge, NJ: Noyes Press, 1981.
- Hanushek, Eric A., and Ludger Woessmann. "The Role of Cognitive Skills in Economic Development?" *Journal of Economic Literature* 46, no. 3 (2008): 607–68.
- Hardie, David W.F. *A Handbook of Modern Breton (Armorican)*. Cardiff: University of Wales Press, 1948.
- Hechter, Michael. *Containing Nationalism*. New York: Oxford University Press, 2000.
- Hobsbawm, Eric. *Nations and Nationalism since 1780: Programme, Myth, Reality*. Cambridge: Cambridge University Press, 1990.
- Kohn, Hans. *The Idea of Nationalism: A Study in Its Origins and Background*. New Brunswick, NJ: Transaction, 2005.
- Laitin, David D. *Identity in Formation: The Russian-Speaking Population in the Near Abroad*. Ithaca: Cornell University Press, 1998.
- . *Nations, States, and Violence*. New York: Oxford University Press, 2007.
- Laitin, David D., Joachim Moortgat, and Amanda Lea Robinson. "Geographical Axes and the Persistence of Cultural Diversity." *Proceedings of the National Academy of Sciences* 109, no. 26 (2012): 10263–68.
- Laitin, David D., and Rajesh Ramachandran. "Language Policy and Human Development." *American Political Science Review* 110, no. 3 (2016): 457–80.
- Landau, Sidney I. *Dictionaries: The Art and Craft of Lexicography*, 2nd ed. New York: Cambridge University Press, 2001.
- Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig, eds. *Ethnologue: Languages of the World*, 17th ed. Dallas, TX: SIL International, 2013. www.ethnologue.com/
- Lindert, Peter H. "Voice and Growth: Was Churchill Right?" *Journal of Economic History* 63, no. 2 (2003): 315–50.
- Liu, Amy H. *Standardizing Diversity: The Political Economy of Language Regimes*. Philadelphia: University of Pennsylvania Press, 2014.
- Martin, Peter. *Samuel Johnson: A Biography*. Cambridge, MA: Harvard University Press, 2008.
- Marvel, Seth A., Travis Martin, Charles R. Doering, et al. "The Small-World Effect Is a Modern Phenomenon." Mimeo, 2013.
- McCloskey, Deirdre N. *Bourgeois Equality: How Ideas, Not Capital or Institutions, Enriched the World*. Chicago: University of Chicago Press, 2016.
- McCusker, John J. "The Demise of Distance: The Business Press and the Origins of the Information Revolution in the Early Modern Atlantic World." *American Historical Review* 110, no. 2 (2005): 295–321.
- McKitterick, David. "The Beginning of Printing." In *The New Cambridge Medieval History*, Vol. 7: c.1415–c.1500, edited by Christopher Allmand, 287–98. Cambridge: Cambridge University Press, 1998.

- McNeill, William H. *The Pursuit of Power: Technology, Armed Force and Society since A.D. 1000*. Chicago: University of Chicago Press, 1982.
- Michalopoulos, Stelios. "The Origins of Ethnolinguistic Diversity." *American Economic Review* 102, no. 4 (2012): 1508–39.
- Minahan, James B. *One Europe, Many Nations: A Historical Dictionary of European National Groups*. Westport, CO: Greenwood Press, 2000.
- Mitchell, Brian R. *International Historical Statistics: Africa, Asia, and Oceania, 1750–2000*, 4th ed. New York: Palgrave Macmillan, 2003a.
- . *International Historical Statistics: Europe, 1750–2000*, 5th ed. New York: Palgrave Macmillan, 2003b.
- Mokyr, Joel. *The Gifts of Athena: Historical Origins of the Knowledge Economy*. Princeton: Princeton University Press, 2002.
- . "Long-term Economic Growth and the History of Technology." In *Handbook of Economic Growth*, Vol. 1B, edited by Philippe Aghion and Steven N. Durlauf, 1113–80. New York: Elsevier, 2005.
- . *A Culture of Growth: The Origins of the Modern Economy*. Princeton: Princeton University Press, 2016.
- Nordhoff, Sebastian, Harald Hammarström, Robert Forkel, et al., eds. *Glottolog 2.0*. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2013. Accessed at <http://glottolog.org>.
- Nunn, Nathan. "Historical Development." In *Handbook of Economic Growth*, Vol. 2A, edited by Philippe Aghion and Steven N. Durlauf, 347–402. New York: Elsevier, 2014.
- Nüssli, Christos. "EurAtlas," 2011. Accessed at <http://www.euratlas.com/about.html>.
- Olsson, Ola, and Christopher Paik. "Long-run Cultural Divergence: Evidence from the Neolithic Revolution." *Journal of Development Economics* 122 (2016): 197–213.
- Ostler, Nicholas. *Empires of the Word: A Language History of the World*. New York: Harper Collins, 2005.
- Pettegree, Andrew. *The Book in the Renaissance*. New Haven: Yale University Press, 2010.
- Pfaff, Steven, and Katie E. Corcoran. "Piety, Power, and the Purse: Religious Economies Theory and Urban Reform in the Holy Roman Empire." *Journal for the Scientific Study of Religion* 51, no. 4 (2012): 757–76.
- Pleiades*. 2015. Accessed at <http://pleiades.stoa.org/home>
- Price, Glanville, ed. *Encyclopedia of the Languages of Europe*. Malden, MA: Blackwell, 1998.
- Reske, Christoph. *Die Buchdrucker des 16. und 17. Jahrhunderts im deutschen Sprachgebiet. Auf der Grundlage des gleichnamigen Werkes von Josef Benzing*. Wiesbaden: Harrassowitz, 2007.
- Riley, Shawn J., Stephen D. DeGloria, and Robert Elliot. "A Terrain Ruggedness Index That Quantifies Topographic Heterogeneity." *Intermountain Journal of Sciences* 6, no. 1–4 (1999): 23–27.
- Rubin, Jared. "Printing and Protestants: An Empirical Test of the Role of Printing in the Reformation." *Review of Economics and Statistics* 96, no. 2 (2014): 270–86.
- Rüegg, Walter, ed. *A History of the University in Europe*, 4 vols. Cambridge: Cambridge University Press, 2011.

- Safran, William, and Jean A. Laponce, eds. *Language, Ethnic Identity and the State*. New York: Routledge, 2005.
- Smith, Anthony D. *The Ethnic Origins of Nations*. Malden, MA: Blackwell, 1986.
- Spolaore, Enrico, and Romain Wacziarg. "The Diffusion of Development." *Quarterly Journal of Economics* 124, no. 2 (2009): 469–529.
- . "How Deep Are the Roots of Economic Development?" *Journal of Economic Literature* 51, no. 2 (2013): 325–69.
- Stasavage, David. "When Distance Mattered: Geographic Scale and the Development of European Representative Assemblies." *American Political Science Review* 104, no. 4 (2010): 625–43.
- . "Was Weber Right? The Role of Urban Autonomy in Europe's Rise." *American Political Science Review* 108, no. 2 (2014): 337–54.
- Steinberg, Sigfrid H. *Five Hundred Years of Printing*, 3rd ed. Middlesex: Penguin Books, 1974.
- Tabellini, Guido. "Institutions and Culture." *Journal of European Economic Association* 6, no. 2–3 (2008): 255–94.
- . "Culture and Institutions: Economic Development in the Regions of Europe." *Journal of European Economic Association* 8, no. 4 (2010): 677–716.
- Talbert, Richard J.A., ed. *Barrington Atlas of the Greek and Roman World*. Princeton: Princeton University Press, 2000.
- Tilly, Charles. "Reflection on the History of European State-Making." In *The Formation of National States in Western Europe*, edited by Charles Tilly, 3–83. Princeton: Princeton University Press, 1975.
- . "War Making and State Making as Organized Crime." In *Bringing the State Back In*, edited by Peter B. Evans, Dietrich Rueschemeyer and Theda Skocpol, 169–91. New York: Cambridge University Press, 1985.
- . *Coercion, Capital, and European States, AD 990–1992*. Malden, MA: Blackwell, 1992.
- . "States and Nationalism in Europe, 1492–1992." *Theory and Society* 23, no. 1 (1994): 131–46.
- Weber, Eugen. *Peasants into Frenchmen: The Modernization of Rural France, 1870–1914*. Stanford: Stanford University Press, 1976.
- Weber, Max. *Economy and Society: An Outline of Interpretive Sociology*. Berkeley: University of California Press, 1978[1922].
- Wimmer, Andreas. *Nationalist Exclusion and Ethnic Conflict: Shadows of Modernity*. New York: Cambridge University Press, 2002.
- Woodberry, Robert D. "The Missionary Roots of Liberal Democracy." *American Political Science Review* 106, no. 2 (2012): 244–74.

Appendix for “Publishing Nations: Technology Acquisition and Language Standardization for European Ethnic Groups”

July 18, 2017

Contents

1	List of 171 Ethnic Groups in the Data Set	A2
2	List of 47 European States in the Data Set	A7
3	Regression Tables	A9
3.1	Cox proportional hazards and logistic regression models used in paper	A9
3.2	Robustness checks I: state as the unit of analysis	A11
3.2.1	State-only sample	A11
3.2.2	Stateless ethnic groups-only sample	A13
3.3	Robustness checks II: The role of human capital	A15
3.4	Instrumental variables (IV) approach: Distance from Mainz as IV	A17

1 List of 171 Ethnic Groups in the Data Set

Table A1

	name	homeland city	printing press	dictionary
1	Abaza	Adyge-Khabl		
2	Abkhaz	Sukhumi		1986
3	Adyge	Maykop		
4	Aguls	Tpig, Dagestan		
5	Ajars	Batumi		
6	Akhvakh	Akhvakhsky district, Dagestan		
7	Alanders	Mariehamm		
8	Albanians	Tirana	1555	1980
9	Alsations	Strasbourg	1460	1899
10	Andalusians	Seville	1477	1941
11	Andis	Khasavyurt, Dagestan		
12	Andorrans	Andorra La Vella		
13	Archis	Arsha-Makhi, Dagestan		1977
14	Armenians	Yerevan	1512	1971
15	Aromanians	Trikala		1963
16	Austrians	Vienna	1461	1963
17	Avars	Kunzakh, Khunzakhsky district, Dagestan		
18	Azeris	Baku	1870	1966
19	Azoreans	Ponta Delgada		
20	Bagulals	Tsumadinsky district		
21	Balkars	Karachayevsky district		1996
22	Bashkorts	Ufa		1993
23	Basques	Vitoria	1545	1989
24	Bats	Tusheti		2003
25	Bavarians	Munich	1482	1827
26	Belarussians	Minsk	1520	1977
27	Bezhetas	Bezhta village, Tsuntinsky district, Dagestan		
28	Bosnians	Sarajevo	1866	1992
29	Botligs	Botlikh village, Botlikhsky district, Dagestan		
30	Bretons	Rennes	1485	1958
31	Budugs	Budug		
32	Bulgarians	Sofia	1566	1895
33	Burgundians	Dijon	1491	
34	Canarians	Santa Cruz de Tenerife		1995
35	Carpatho-Rusyns	Uzhgorod	1572	

continued on next page

continued from previous page

	name	homeland city	printing press	dictionary
36	Catalans	Barcelona	1475	1930
37	Chamalals	Agvali village, Tsumadinsky district, Dagestan		
38	Chavash	Cheboksary		1928
39	Chechens	Grozny		1961
40	Cherkess	Cherkessk		1956
41	Cornish	Truno	1700	
42	Corsicans	Ajaccio	1769	
43	Crimean Tatars	Bakhchysarai		
44	Croats	Zagreb	1483	1901
45	Csangos	Miercurea Ciuc	1533	1936
46	Czechs	Prague	1487	1935
47	Danes	Copenhagen	1490	1918
48	Dargins	Makhachkala, Dagestan		2005
49	Didos	Kidero village, Tsuntinsky district, Dagestan		
50	Don Cossacks	Novocherkassk		1975
51	Dutch	Amsterdam	1477	1864
52	English	London	1476	1755
53	Estonians	Tallinn	1535	1918
54	Faroese	Torshavn		1966
55	Finns	Helsinki	1542	1953
56	Flemish	Brussels	1475	1873
57	French	Paris	1470	1858
58	Frisians	Leeuwarden	1630	1874
59	Friulis	Udine	1484	1985
60	Gagauz	Comrat		
61	Galicians	Santiago de Compostela	1483	1986
62	Georgians	Tbilisi	1709	1884
63	Germans	Berlin	1540	1852
64	Gibraltarians	Gibraltar	1878	
65	Ginugs	Tsuntinsky district, Dagestan		
66	Godoberis	Godoberi village, Botlikhsky district, Dagestan		
67	Greeks	Athens	1759	1936
68	Guernseians	St. Peter Port		
69	Hungarians	Budapest	1473	1890
70	Icelanders	Reykjavik	1540	1912
71	Ingrians	Saint Petersburg	1711	1953
72	Ingush	Nazran		
73	Irish	Dublin	1550	1913
74	Istrians	Pulj	1905	1979
75	Italians	Rome	1467	1861

continued on next page

continued from previous page

	name	homeland city	printing press	dictionary
76	Jerseyites	St. Helier		
77	Kabards	Nalchik		
78	Kalmyks	Elista		
79	Karachais	Mikoyan Shakhar		1996
80	Karatas	Karata village, Akhvakhsky district, Dagestan		
81	Karels	Petrozavodsk	1930	1968
82	Kashubians	Gdansk	1499	1994
83	Khinalugs	Quba		
84	Khunzuls	Nakhada village, Tsuntinsky district, Dagestan		
85	Khwarshia	Khvarshi village, Tsumadinsky district, Dagestan		
86	Kists	Kakheti		
87	Komis	Syktyvkar		1959
88	Kosovars	Pristina		
89	Kryts	Jek		
90	Kuban Cossacks	Krasnodar		
91	Kумыks	Buynaksk, Dagestan		
92	Ladins	Belluno	1475	
93	Laks	Vachi, Dagestan		
94	Latvians	Riga	1525	1972
95	Lezgins	Mamash, Dagestan		
96	Liechtensteiners	Vaduz		
97	Ligurians	Genoa	1472	1985
98	Lithuanians	Vilnius	1525	1956
99	Livonians	Ventspils	1542	1938
100	Lombards	Milan	1471	1835
101	Luxembourgers	Luxembourg City	1598	1950
102	Macedonians	Skopje	1840	1961
103	Madeirans	Funchal		
104	Maltese	Valletta	1642	1987
105	Manx	Douglas	1707	1835
106	Maris	Yoshkar-Ola	1767	1990
107	Megleno-Romanians	Kilkis	1512	1963
108	Meskhtekians	Akhaltikhe		
109	Moldovans	Chisinau	1642	1977
110	Monegasques	Monaco-Ville	1795	

continued on next page

continued from previous page

	name	homeland city	printing press	dictionary
111	Montenegrins	Podgorica	1494	
112	Moravians	Brno	1486	1982
113	Mordvins	Saransk		1973
114	Mountain Jews	Majalis, Dagestan		
115	Nenets	Naryan-Mar		
116	Nogais	Kizlyar, Dagestan		
117	Normans	Rouen	1485	
118	Northumbrians	Newcastle	1646	
119	Norwegians	Oslo	1643	1930
120	Occitans	Toulouse	1476	1875
121	Ossetians	Vladikavkaz		1958
122	Piedmontese	Turin	1474	1983
123	Poles	Warsaw	1522	1807
124	Pomaks	Edirne	1554	
125	Portuguese	Lisbon	1489	1852
126	Rhinelanders	Cologne	1464	1928
127	Romands	Geneva	1478	1926
128	Romanians	Bucharest	1708	1907
129	Romansh	Chur	1557	1939
130	Romas	Soroca		
131	Russians	Moscow	1564	1891
132	Rutuls	Rutul, Dagestan		
133	Samis	Kautokeino	1649	1939
134	Sanjakis	Novi Pazar		
135	San Marinese	San Marino		
136	Sards	Cagliari	1493	
137	Savoyards	Chambéry	1484	1969
138	Saxons	Dresden	1524	1953
139	Scanians	Malmo	1528	
140	Scots	Edinburgh	1508	1931
141	Serbs	Belgrade	1552	1818
142	Sicilians	Palermo	1478	1977
143	Silesians	Wroclaw	1475	1962
144	Slovaks	Bratislava	1500	1959
145	Slovenes	Ljubljana	1575	1976
146	Sorbs	Bautzen	1555	1978
147	Spaniards	Madrid	1566	1886
148	Swabians	Stuttgart	1486	1901
149	Swedes	Stockholm	1483	1893
150	Swiss-Germans	Zürich	1479	1881

continued on next page

continued from previous page

	name	homeland city	printing press	dictionary
151	Szeklars	Cluj-Napoca	1578	1975
152	Tabasarans	Khiv, Dagestan		
153	Talysh	Astara		
154	Tats	Shabran		
155	Tatars	Kazan	1804	1969
156	Terek Cossacks	Stavropol		
157	Tindis	Tindi village, Tsumadinsky district, Dagestan		
158	Tsakhurs	Balakan, Dagestan		
159	Turks	Ankara	1726	1963
160	Tyroleans	Innsbruck	1547	1866
161	Udis	Qabala district		
162	Udmurts	Izhhevsk		1890
163	Ukrainians	Kiev	1616	1886
164	Venetians	Venice	1469	1829
165	Veps	Sheltozero district		1972
166	Volga Germans	Saratov		1974
167	Vorarlbergers	Bregenz	1658	1955
168	Waldensians	Torre Pellice	1479	1973
169	Walloons	Namur		1927
170	Welsh	Cardiff	1546	1887
171	Zetlanders	Lerwick		

2 List of 47 European States in the Data Set

This table provides a list of 47 European states in my dataset, which describes the earliest date of the printing press, the name of the city where the first printing press was set up (whenever the name is available), the language with the largest population, and the date of the vernacular dictionary. The *Language* column denotes the official languages; for countries with multiple official languages, the language with the largest population is listed on the first row and those with fewer populations are on the second row.

The data is sorted by the date of print acquisition.

Table A2

	country	printing press	city	dictionary	language
1	Germany	1450	Mainz	1852	German (standard)
2	Austria	1461	Vienna	1963	German (regional)
3	Italy	1464	Subiaco	1861	Italian
4	Netherlands	1465	Haarlem	1864	Dutch
5	France	1470	Paris	1858	French
6	Switzerland	1470	Basel, Beromünster	1881	Swiss-German, French, Italian, Romansh
7	Belgium	1473	Bruges	1927	French (regional), Dutch
8	Hungary	1473	Buda (Budapest)	1890	Hungarian
9	Spain	1473	Valencia, Barcelona	1886	Spanish
10	Poland	1474	Kraków (Cracow)	1807	Polish
11	Czech Republic	1476	Plzeň (Pilsen)	1935	Czech
12	United Kingdom	1476	London	1755	English
13	Denmark	1482	Odense	1918	Danish
14	Croatia	1483	Zagreb	1901	Croatian
15	Sweden	1483	Stockholm	1893	Swedish
16	Portugal	1487	Faro	1852	Portuguese
17	Turkey	1493	Istanbul	1963	Turkish
18	Montenegro	1494	Podgorica		Montenegrin
19	Slovakia	1500	Bratislava	1959	Slovak
20	Romania	1508	Târgoviște	1907	Romanian (standard)
21	Armenia	1512	Yerevan	1971	Armenian
22	Greece	1512	Salonica (Thessaloniki)	1936	Greek
23	Belarus	1520		1977	Belarusian, Russian
24	Lithuania	1525	Vilnius	1956	Lithuanian
25	Iceland	1530	Hollar	1912	Icelandic

continued on next page

continued from previous page

	name	printing press	city	dictionary	language
26	Bosnia and Herzegovina	1531	Goražde	1992	Bosnian
27	Ireland	1550	Dublin	1913	Irish, English
28	Serbia	1552	Belgrade	1818	Serbian
29	Albania	1555	Tirana	1980	Albanian
30	Russia	1564	Moscow	1891	Russian
31	Ukraine	1574	Lviv (Lvov)	1886	Ukrainian
32	Slovenia	1575	Ljubljana	1976	Slovene
33	Latvia	1587	Riga	1972	Latvian
34	Luxembourg	1598	Luxembourg	1950	Luxemburgish, French, German
35	Estonia	1630	Tallinn	1918	Estonian
36	Finland	1642	Turku	1953	Finnish, Swedish
37	Malta	1642		1987	Maltese
38	Moldova	1642	Iași	1977	Romanian (regional)
39	Norway	1643	Olso	1930	Norwegian
40	Georgia	1709	Tbilisi	1884	Georgian
41	Monaco	1795			Monégasque
42	Bulgaria	1828	Samokov	1895	Bulgarian
43	Macedonia	1840		1961	Macedonian
44	Azerbaijan	1870	Baku	1966	Azeri
45	Andorra				
46	San Marino				
47	Liechtenstein				

3 Regression Tables

3.1 Cox proportional hazards and logistic regression models used in paper

Table A3 reports the complete regression estimates of Table 3 of the main body.

Table A3: Regression outputs of the printing press' impact on language standardization.

model	bivariate		demand side		supply side		fully specified	
	Cox (1)	logit (2)	Cox (3)	logit (4)	Cox (5)	logit (6)	Cox (7)	logit (8)
Printing press	2.159*** (0.269)	2.933*** (0.342)	1.600*** (0.356)	2.178*** (0.456)	1.455*** (0.354)	1.932*** (0.503)	1.450*** (0.339)	2.083*** (0.556)
University			0.0004*** (0.0001)	0.001*** (0.0002)	0.0004*** (0.0001)	0.001*** (0.0003)	0.0004*** (0.0001)	0.001*** (0.0002)
Bishop			-0.000002 (0.0001)	-0.0001 (0.0002)	0.00003 (0.0001)	0.00004 (0.0002)	-0.00002 (0.0001)	-0.00002 (0.0003)
Log urban potential			0.114 (0.226)	0.098 (0.365)	-0.002 (0.607)	-0.307 (0.793)	0.186 (0.636)	-0.276 (0.821)
Log distance to Wittenberg/Zürich					0.0001 (0.0003)	-0.0001 (0.001)	0.0004 (0.0003)	0.0002 (0.001)
War frequency					-0.031** (0.014)	-0.043*** (0.016)	-0.030** (0.013)	-0.048*** (0.018)
Elevation					-0.0001 (0.0004)	-0.00002 (0.001)	0.0001 (0.0004)	0.0005 (0.001)
Terrain ruggedness					-0.003** (0.001)	-0.004** (0.002)	-0.002** (0.001)	-0.004** (0.002)
Hub Roman road					0.156 (0.296)	0.494 (0.634)	0.129 (0.337)	0.431 (0.7)
Roman road					-0.009 (0.247)	-0.023 (0.467)	0.081 (0.304)	0.151 (0.603)
Oceanic port					-0.415** (0.212)	-0.712* (0.411)	-0.477* (0.26)	-0.721 (0.539)
Island					-0.057 (0.475)	0.185 (0.85)	-0.038 (0.52)	0.2 (0.965)
Ottoman Empire							-0.621 (0.812)	-0.822 (0.959)
Russian Empire							0.243 (0.237)	0.354 (0.473)
Habsburg Empire							-0.075 (0.33)	-0.252 (0.704)
Orthodox							0.279 (0.284)	0.725 (0.614)
Protestant							0.383 (0.235)	0.613 (0.474)
Islam							-0.939** (0.446)	-1.072 (0.694)

continued on next page

continued from previous page

model	bivariate		demand side		supply side		fully specified	
	Cox (1)	logit (2)	Cox (3)	logit (4)	Cox (5)	logit (6)	Cox (7)	logit (8)
Northern Europe							-0.246 (0.35)	-0.371 (0.719)
Southern Europe							0.017 (0.317)	-0.264 (0.695)
Eastern Europe							-0.194 (0.363)	-0.39 (0.768)
Western Asia							-0.621 (0.595)	-1.207 (1.043)
t		0.024 (0.017)		0.021 (0.017)		0.04 (0.029)		0.04 (0.029)
t^2		0.00001 (0.00002)		0.00001 (0.00002)		-0.00001 (0.00004)		-0.00001 (0.00004)
Intercept		-16.151*** (3.697)		-16.159*** (3.767)		-19.244*** (6.554)		-19.811*** (6.452)
AIC	1664.72	516.44	1650.75	494.78	1645.65	484.26	1645.01	480.83
Observations	2,052	2,052	2,052	2,052	2,052	2,052	2,052	2,052

Note: Robust standard errors clustered by ethnic groups for all models. *Western Europe* is used as the reference category for region fixed effects and thus omitted. *Catholic* is used as the reference category for religion fixed effects and thus omitted. *** denote $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

3.2 Robustness checks I: state as the unit of analysis

3.2.1 State-only sample

Table A4 reports the complete output of the abridged Cox proportional hazards models shown in Table 4 of the main body. In this series of robustness checks, I use states as the unit of analysis. “List” refers to a set of alternative dates of dictionary publications for countries with multiple official languages in the dependent variable.

Table A4: Cox proportional hazards model outputs of the printing press’ impact on language standardization.

model	List 1 (1)	List 2 (2)	List 3 (3)	List 4 (4)
Printing press	14.711*** (0.98)	14.984*** (0.92)	14.725*** (1.1)	14.814*** (0.97)
First year of printing press	-0.002 (0.001)	-0.001 (0.001)	-0.002* (0.001)	-0.002 (0.001)
University	0.0002 (0.0002)	0.0003 (0.0002)	0.0001 (0.0002)	0.0002 (0.0002)
Bishop	-0.0003** (0.0002)	-0.0002 (0.0002)	-0.0001 (0.0002)	-0.0002 (0.0002)
Log urban potential	0.932 (0.595)	1.335** (0.532)	1.254** (0.58)	0.937 (0.62)
Log distance to Wittenberg/Zürich	0.0007* (0.0004)	0.001* (0.0003)	0.001** (0.0004)	0.001* (0.0004)
War frequency	-0.003 (0.013)	-0.001 (0.012)	-0.002 (0.013)	-0.006 (0.013)
Elevation	-0.002*** (0.0005)	-0.001 (0.0006)	-0.002*** (0.0005)	-0.002*** (0.0005)
Terrain ruggedness	-0.005* (0.003)	-0.006** (0.003)	-0.005* (0.003)	-0.005* (0.003)
Hub Roman road	0.336 (0.364)	0.468 (0.399)	0.547 (0.427)	0.462 (0.377)
Roman road	0.402 (0.324)	-0.028 (0.338)	0.234 (0.38)	0.378 (0.324)
Oceanic port	-0.8** (0.337)	-0.59* (0.351)	-0.886*** (0.344)	-0.805*** (0.336)
Island	0.115 (0.258)	0.268 (0.271)	0.160 (0.262)	0.118 (0.25)

continued on next page

continued from previous page

model	List 1 (1)	List 2 (2)	List 3 (3)	List 4 (4)
Ottoman Empire	-1.822* (1.1)	-2.127** (1.06)	-1.949* (1.09)	-1.702 (1.1)
Russian Empire	-0.244 (0.289)	-0.054 (0.251)	-0.225 (0.295)	-0.216 (0.298)
Habsburg Empire	-0.642 (0.441)	-0.925** (0.455)	-0.863* (0.503)	-0.641 (0.45)
Orthodox	-0.095 (0.462)	0.0002 (0.453)	-0.164 (0.535)	-0.092 (0.472)
Protestant	1.250*** (0.303)	0.268 (0.31)	0.885*** (0.316)	1.136*** (0.317)
Islam	-0.371 (1.163)	-0.267 (0.948)	0.020 (1.051)	-0.389 (1.173)
Northern Europe	-0.378 (0.303)	0.110 (0.311)	-0.343 (0.312)	-0.259 (0.307)
Southern Europe	0.685* (0.359)	0.537 (0.395)	0.629 (0.4)	0.780** (0.374)
Eastern Europe	1.105** (0.5)	0.640 (0.526)	0.758 (0.593)	1.074** (0.513)
Western Asia	1.592* (0.963)	0.757 (0.965)	1.076 (0.958)	1.646* (0.94)
AIC	538.29	596.87	558.7	533.87
Observations	564	564	564	564

Note: Robust standard errors clustered by country for all models. *Western Europe* and *Catholic* are used as the reference category, respectively, for geographic region fixed effects and for religion fixed effects so they are omitted. *** denote $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

3.2.2 Stateless ethnic groups-only sample

Table A5 documents Cox regression estimates using stateless ethnic groups ($n = 124$). This is reported in Column 5 of Table 4 in the main body.

Table A5: Regression outputs of the printing press' impact on language standardization.

model	Cox (1)
Printing press	4.915** (2.293)
Print year	-0.002* (0.001)
University	0.0002 (0.0002)
Bishop	-0.00005 (0.0002)
Log urban potential	0.700 (0.967)
Log distance to Wittenberg/Zürich	0.001 (0.0005)
War frequency	-0.032 (0.02)
Elevation	0.0005 (0.0006)
Terrain ruggedness	-0.004** (0.002)
Hub Roman road	-0.254 (0.396)
Roman road	0.021 (0.369)
Oceanic port	-0.434 (0.395)
Island	-0.438 (0.966)
Ottoman Empire	-15.591*** (0.875)
Russian Empire	0.979*** (0.297)
Habsburg Empire	0.543 (0.571)

continued on next page

continued from previous page

model	Cox (1)
Orthodox	0.202 (0.403)
Protestant	0.469 (0.297)
Islam	-1.295** (0.571)
Northern Europe	-0.413 (0.537)
Southern Europe	0.136 (0.48)
Eastern Europe	-1.079*** (0.412)
Western Asia	-2.881** (1.328)
AIC	815.41
Observations	1,488

Note: Robust standard errors clustered by country for all models. *Western Europe* and *Catholic* are used as the reference category, respectively, for geographic region fixed effects and for religion fixed effects and are omitted. *** denote $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

3.3 Robustness checks II: The role of human capital

Table A6 documents the full regression estimates of Table 5 of the main body.

Table A6: Regression outputs of the printing press' impact on language standardization.

model	Cox (1)	logit (2)	Cox (3)	logit (4)
Printing press			1.394*** (0.342)	2.023*** (0.558)
Vernacular Bible	0.531** (0.251)	0.745** (0.356)	0.461* (0.257)	0.676* (0.365)
University	0.0005*** (0.0001)	0.001*** (0.0002)	0.0003** (0.0001)	0.001*** (0.0003)
Bishop	0.00006 (0.0001)	0.0001 (0.0003)	-0.00004 (0.0001)	-0.0001 (0.0003)
Log urban potential	-0.157 (0.613)	-0.554 (0.805)	0.168 (0.616)	-0.259 (0.818)
Log distance to Wittenberg/Zürich	-0.0001 (0.0003)	-0.0004 (0.001)	0.0004 (0.0003)	0.0002 (0.001)
War frequency	-0.026** (0.013)	-0.038** (0.016)	-0.03** (0.013)	-0.048*** (0.018)
Elevation	0.0003 (0.0004)	0.001 (0.001)	0.0002 (0.0004)	0.001 (0.001)
Terrain ruggedness	-0.003*** (0.001)	-0.005** (0.002)	-0.003** (0.001)	-0.004** (0.002)
Hub Roman road	0.217 (0.341)	0.531 (0.693)	0.054 (0.336)	0.267 (0.712)
Roman road	0.007 (0.29)	0.082 (0.582)	0.107 (0.31)	0.252 (0.62)
Oceanic port	-0.213 (0.252)	-0.317 (0.487)	-0.436* (0.263)	-0.658 (0.541)
Island	-0.36 (0.507)	-0.374 (0.888)	-0.115 (0.511)	0.063 (0.949)
Ottoman Empire	-0.666 (0.792)	-0.914 (0.907)	-0.589 (0.792)	-0.834 (0.892)
Russian Empire	0.018 (0.252)	-0.046 (0.491)	0.173 (0.235)	0.244 (0.472)
Habsburg Empire	0.075 (0.341)	0.007 (0.646)	-0.088 (0.337)	-0.265 (0.7)

continued on next page

continued from previous page

model	Cox (1)	logit (2)	Cox (3)	logit (4)
Orthodox	0.467 (0.291)	0.963 (0.588)	0.286 (0.286)	0.786 (0.602)
Protestant	0.337 (0.248)	0.561 (0.494)	0.345 (0.236)	0.527 (0.482)
Islam	-0.913** (0.454)	-1.031 (0.661)	-0.870* (0.448)	-1.016 (0.679)
Northern Europe	-0.179 (0.346)	-0.259 (0.699)	-0.188 (0.347)	-0.304 (0.717)
Southern Europe	0.006 (0.314)	-0.271 (0.700)	0.001 (0.313)	-0.363 (0.72)
Eastern Europe	-0.190 (0.376)	-0.343 (0.747)	-0.102 (0.371)	-0.281 (0.753)
Western Asia	-0.374 (0.658)	-0.763 (1.025)	-0.650 (0.594)	-1.296 (1.039)
t		0.044 (0.03)		0.041 (0.029)
t^2		-0.00001 (0.00004)		-0.00001 (0.00004)
Intercept		-19.325*** (6.777)		-20.064*** (6.476)
AIC	1660.38	495.57	1642.95	478.43
Observations	2,052	2,052	2,052	2,052

Note: Robust standard errors clustered by ethnic groups for all models. *Western Europe* is used as the reference category for region fixed effects and thus omitted. *Catholic* is used as the reference category for religion fixed effects and thus omitted. *** denote $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

3.4 Instrumental variables (IV) approach: Distance from Mainz as IV

Table A7 reports the full regression estimates of Table 6 of the paper.

Table A7: IV probit regression outputs of the printing press' impact on language standardization.

dependent variable N cumulative dictionary	First stage	Second stage			
	printing press (1)	language standardization			
		by 1850 n = 7 (2)	by 1900 n = 27 (3)	by 1950 n = 48 (4)	by 2000 n = 104 (5)
Log distance to Mainz	-0.159*** (0.060)				
Printing press		11.822 (37.038)	4.639 (3.051)	4.741** (2.284)	4.018** (1.828)
Print adoption year	0.0003*** (0.00004)	0.001 (0.006)	-0.001 (0.001)	-0.001 (0.001)	-0.0004 (0.001)
University	0.0001*** (0.00003)	0.0001 (0.005)	0.0002 (0.0003)	0.0001 (0.0003)	0.0002 (0.0002)
Bishop	0.00001 (0.00003)	-0.001 (0.004)	-0.0001 (0.0002)	-0.0002 (0.0002)	-0.00003 (0.0001)
Log urban potential	-0.004 (0.073)	7.111** (3.010)	2.221*** (0.217)	1.280*** (0.099)	1.275*** (0.074)
Log distance to Wittenberg/Zürich		-0.005 (0.005)	0.001** (0.001)	0.001*** (0.0003)	0.001*** (0.0003)
War frequency	-0.002 (0.003)	0.007 (0.092)	-0.011 (0.022)	-0.022* (0.013)	-0.025*** (0.009)
Elevation	0.0001** (0.00003)	0.005 (0.011)	-0.0002 (0.001)	-0.0004 (0.001)	-0.0001 (0.0004)
Terrain ruggedness	-0.0002** (0.0001)	-0.044 (0.070)	-0.002 (0.003)	-0.001 (0.002)	-0.001 (0.001)
Hub Roman road	0.131** (0.052)	-6.338 (4.421)	-0.458 (0.596)	-0.480 (0.456)	-0.315 (0.390)
Roman road	-0.008 (0.040)	1.358 (3.645)	0.473 (0.576)	0.268 (0.449)	0.141 (0.349)
Oceanic port	0.041 (0.043)	0.444 (3.262)	-0.380 (0.459)	-0.527 (0.366)	-0.538* (0.286)
Island	-0.005 (0.053)	6.790*** (2.019)	1.158** (0.502)	0.452 (0.429)	0.392 (0.417)

continued on next page

continued from previous page

dependent variable N cumulative dictionary	First stage	Second stage			
	printing press	language standardization			
		by 1850 n = 7	by 1900 n = 27	by 1950 n = 48	by 2000 n = 104
	(1)	(2)	(3)	(4)	(5)
Ottoman Empire	-0.049 (0.045)	4.178** (1.643)	-0.220 (0.611)	-0.151 (0.480)	-0.322 (0.467)
Russian Empire	0.020 (0.026)	-2.738*** (0.766)	0.253 (0.526)	0.262 (0.351)	0.320 (0.275)
Habsburg Empire	0.118** (0.046)	-3.027** (1.300)	-0.673 (0.777)	-0.668 (0.596)	-0.590 (0.445)
Orthodox	-0.144** (0.060)	2.110 (3.658)	0.865 (0.594)	0.544 (0.452)	0.526 (0.352)
Catholic/Protestant	-0.213*** (0.043)	-0.390 (5.030)	-0.390 (0.697)	-0.454 (0.468)	-0.269 (0.372)
Islam	-0.124** (0.056)	0.158 (4.829)	-4.899*** (0.984)	-5.652*** (0.609)	-0.680* (0.386)
Northern Europe	-0.001 (0.062)	6.131*** (1.878)	0.020 (0.480)	0.270 (0.400)	0.109 (0.387)
Southern Europe	0.156** (0.067)	5.603*** (2.136)	0.102 (0.549)	-0.297 (0.395)	-0.058 (0.315)
Eastern Europe	0.060 (0.070)	9.491** (4.449)	0.614 (0.846)	0.044 (0.591)	0.017 (0.444)
Western Asia	0.071 (0.078)	11.974* (6.995)	1.308 (1.532)	0.198 (1.097)	-0.566 (0.622)
Intercept	1.28*** (0.413)	-28.929 (18.163)	-9.522*** (1.138)	-6.967*** (0.749)	-6.992*** (0.598)
Observations	2,052	1,539	1,710	1,881	2,052
p-value, Wald exogeneity test		0.75	0.12	0.038*	0.028*
F-statistic on weak instrument	24.99				

Note: Robust standard errors clustered by ethnic groups. IV probit estimation: first stage is OLS, second stage is probit, regressed on predicted values from the first stage. In the first stage, *log urban potential* and *war frequency* are for the fourteenth centuries. *Western Europe* is used as the reference category for region fixed effects and thus omitted. In the second stage, *Catholic* is used as the reference category for religion fixed effects and thus omitted. *** denote $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.