Reconstructing History: Using Language to Estimate Religious Spread

Arthur Blouin and Julian Dyer

We introduce a data-driven approach to use language to reconstruct history, and apply the methodology to estimate the geographic origins of religious spread. To validate the approach, we use language data to estimate origins of Islam and Buddhism to within 500km of their true (and uncontested) origins. We then apply the methodology to the more complex (and contested) cases of Christianity, Judaism, and Hinduism. We show that language-based estimates, in these cases, are significantly more aligned with the origin of scripture than with the origin of the religion.

Reconstructing the vast share of human history that remains unrecorded has long been a crucial, but challenging, task for historians. This task is made even more difficult when historians study contexts with incomplete survival of historical records, or from places and eras that did not keep easily interpreted records in the first place. The main approach to deal with this issue is to study archaeological evidence, which—while reliable—is costly and heavily localized. Another method of reconstructing history has been to consider the information contained in a society's language. This approach has been prevalent for centuries, having been touted at least as far back as 1765 as the one "that serve[s] best for determining the origin of peoples" (Leibniz 1996 translation, p. 285). Nearly 200 years later, using language to reconstruct history was called "one of the triumphs of nineteenth-century science" (Bloomfield 1939, p. 124).

However, while this approach remains heavily relied on today, its use is controversial, and its validity is vigorously debated. In fact, it has faced skepticism over "arbitrary and unrigorous methods" (Coleman 1988, p. 450), concerns that "semantic reconstruction lacks rigor" (Diebold 1994,

The Journal of Economic History. doi: 10.1017/S0022050725100867

Arthur Blouin is Associate Professor, University of Toronto, 150 St. George Ave., Rm 305, Toronto, Ontario, M5S 3G7. E-mail: a.blouin@utoronto.ca (corresponding author). Julian Dyer is Lecturer, University of Exeter, Rennes Drive, Exeter, EX4 4PU. E-mail: j.dyer3@exeter.ac.uk.

We acknowledge financial support from SSHRC and the Connaught fund. We thank Sascha Becker, Alberto Bisin, Andrew Oswald, Jared Rubin, Elliott Ash, and seminar participants at NYU and UBC for helpful comments. Elissa Chrapko, Kamilah Lewis, and Zi Wei Low all provided outstanding research assistance. We also thank the editor Bishnupriya Gupta and two anonymous referees for their very helpful feedback, which greatly improved the article.

p. 2909), and that it is "notoriously subject to individual interpretation" (Lehmann 1968, p. 404). That said, if a rigorous empirical approach to using semantic analysis to reconstruct history was available, it could open new opportunities for scholars to study unrecorded history.

The main goals of this article are twofold. First, to provide a proof-of-concept assessment of whether the practice of using linguistic clues to reconstruct history can be accurately applied in an objective and rigorous manner. Second, if language can help to reconstruct history, we hope to shed some light on what parts of history language can help to identify. To accomplish these goals, we start by constructing a database with a global scope that identifies *loanwords* and their source languages using machine-learning techniques. Loanwords are words that, at some point in history, have been adopted from another society.[1] Using the loanwords data, we construct topic-specific language-networks, and identify the most influential members of these networks.

We use the loanwords data to explore the geographic origins of the spread of the world's five major religions.[2] Religion is an apt application for our purposes because there are religious words in essentially all languages; religion is an important feature of the global landscape (Pascali 2016; Valencia Caicedo 2019; Becker and Pascali 2019; Valencia Caicedo, Dohmen, and Pondorfer 2021; Becker and Pfaff 2022); and the potential origins of spread of each of the five major religions have been thoroughly studied.

We start by validating our methodology. To do so, we focus on the origins of Buddhism and Islam and demonstrate that our approach can accurately estimate the geographic locations where the global spread of these religions originated. These religions have well-known and uncontested origins, allowing us to provide evidence that our methodology successfully identifies the correct locations.[3] This validation exercise suggests that loanwords do hold significant informational value. The historical account and our estimated origin of spread for Buddhism and Islam are each less than 500km away from each other (and, on average, about 370km away).[4] However, when linguistic information is excluded,

[1] Loanwords are distinguished from cognates, which are words with common linguistic ancestry, and neologisms, which are newly innovated words.

[2] These are: (1) Buddhism; (2) Hinduism; (3) Islam; (4) Judaism; (5) Christianity.

[3] There is a large body of work using complementary applications of non-etymological forms of historical information in language to answer other questions, such as Yu and Huangfu (2019), Baledent, Hiebel, and Lejeune (2020), and Assael et al. (2022), to name a few.

[4] We calibrate our estimates using both Islam and Buddhism to avoid a mechanical estimate of either one. While the Buddhism estimate is slightly closer when we calibrate using Buddhism, and likewise for Islam, the estimates are still only 399km off on average if we rely just on the estimate of Buddhism calibrated using Islam and the Islam estimate calibrated using Buddhism.

the estimates are about 1,300km away. This suggests that methods that draw on etymology to make historical inferences are empirically valid.

After showing that language can help to trace the historical origins of religion, we apply the methodological approach to explore the more complex cases of Judaism, Christianity, and Hinduism, where there is greater debate and uncertainty surrounding their origins. Much of this uncertainty stems from the fact that the global spread may have originated from canonical religious texts, or *scripture* (Rubin 2014), rather than from the early adherents to a particular religion.[5] Accordingly, language-based estimates could reflect origin locations of words spread orally (via preaching) or in writing (via scripture). Understanding this nuance could be crucial for future applications that rely on language to reconstruct history, since these locations are often very different from one another.

The spread of Christianity, for example, could be seen as emanating from Greece, where the gospel was preached by Paul; Alexandria, where the first canonical Christian scripture was written; Constantinople, where the first Christian state was centered; or Jerusalem, where Jesus was born. For Judaism, the origin could be Jerusalem, or near Babylon, where Jews were exiled and first wrote scripture to preserve Jewish traditions. In the case of Hinduism, theories suggest an origin of the scripture in the Indus Valley or the Bactria–Margiana Archaeological Complex (BMAC) region, while the first practicing Hindus are often thought to have originated from the Pontic Steppe.

Thus, for each of Christianity, Judaism, and Hinduism, the geographic origins of scripture are different from the origins of the religion itself or of sacred religious figures. In each of these three cases, we find that the estimates are much nearer to the origin of the scripture than to the origin of the religion itself. This proof-of-concept evidence from religious spread suggests that methods based on linguistic change may primarily identify the textual or canonical origin of a historical phenomenon rather than the geographic origin of the phenomenon itself.

So, while language does appear to contain historically relevant information, some caution is certainly warranted. As noted previously, we should be careful about how to interpret language-based location estimates. Because of this, the methodological approach should not be viewed as a substitute for traditional historical analysis, nor is it suitable as such. Even beyond issues related to interpretation and context, as one might expect in a completely automated approach that does not incorporate historical source information, the estimates are relatively noisy and much

[5] Henceforth, we use scripture to reference the canonical sacred texts of any religion.

less precise than traditional historical analysis. Accordingly, the specific implementation of the approach we investigate in this article may be less helpful for supporting traditional historical analysis when written records are plentiful than for situations where there is no historical scholarship or where the historical scholarship that exists is heavily contested.[6] Second, we automate the entire process because it helps to "tie our hands," which, from an empirical proof-of-concept perspective, is desirable, especially in light of the typical critiques that linguistic historical reconstruction is too "subject to individual interpretation" (Lehmann 1968, p. 404). However, there are trade-offs with this approach. For instance, it seems likely that integrating additional historical facts could greatly improve the accuracy of the approach; however, doing so is beyond the scope of our analysis.

Our main contribution to the literature is to highlight that language can be helpful in reconstructing history when primary source data is missing. There is already a literature that aims to estimate the historical origins of various phenomena. For example, Nunn and Wantchekon (2011) demonstrate that slave trade hubs were the historical origin of mistrust in Africa. In the same vein, Lowes and Montero (2021) highlight the colonial roots of mistrust in medicine in West and Central Africa. In these cases, the object of historical reconstruction is a cultural feature of a society.[7] In our case, we identify the geographic origins of the diffusion of ideas. While we consider the case of religion as a demonstration of the approach, it seems possible that a similar approach could be used to study the spread of various under-documented historical phenomena that may be of interest to economists. This includes a wide range of topics, from the spread of markets to the diffusion of various technologies to various cultural attributes, as in both Nunn and Wantchekon (2011) and Lowes and Montero (2021).

A second contribution to the literature relates to our construction of novel data using machine-learning methods. This approach, summarized in Abramitzky et al. (2021) and Bailey et al. (2020), has recently become more prevalent in economic history. For example, both Feigenbaum (2016) and Price et al. (2021) develop and validate the use of machine-learning methods to link individuals across administrative data sets to generate long-run historical panels. These methods, in addition to overlapping in their aim to construct better data for the purpose of research

[6] We believe that this approach, given that it does not require written sources, will provide the greatest benefit where such written records are unavailable. This may include applications crucial to the study of long-run economic development. This could include the emergence and spread of technology, states, and other social institutions in less-developed regions of the world, though this is beyond the scope of this paper.

[7] And many others; see, for instance, Alesina and Giuliano (2015) for a review of the literature.

in economic history, are also similar in methodology. Just as in our application, Feigenbaum (2016) and Price et al. (2021) rely on orthographic similarity measures in their matching algorithms. In our case, we augment this information with other linguistic features, such as phonetic similarity, which improves performance in our case and may therefore have more general applications in the records-matching literature.

HISTORICAL BACKGROUNDS

Loanwords as Historical Artefacts

This article explores whether the information contained in the etymology, or origin, of words in a society's vocabulary contains information about the evolution of important historical phenomena. This builds on the idea that words themselves contain important information about a group's past experiences. The idea that there is informational content in language is not new. There is a long tradition in linguistics examining how changes in the words a society uses relate to their history and evolution.

> *A community is known by the language it keeps, and its words chronicle the times. Every aspect of the life of a people is reflected in the words they use to talk about themselves and the world around them. As their world changes—through invention, discovery, revolution, evolution, or personal transformation—so does their language. Like the growth rings of a tree, our vocabulary bears witness to our past. (Algeo 1993)*

One aim of this article is to understand whether a linguistic measure of the intensity of cross-societal influence related to a given phenomenon, in our case religion, is useful for tracing the origins of these phenomena. Since we are interested in understanding the nature of cross-societal influence in the religious domain, we follow standard practice to interpret borrowed words related to a given topic as an indicator of influence concerning that topic.

Consider, for instance, the Lakhmid kingdom, which comprised parts of what is now Saudi Arabia (circa 300–600 CE), and for which it has been notoriously difficult to reconstruct a history. Loanwords have helped to trace the roots of their formal institutions: "[…] the Lakhmids, while remaining Arab, inevitably picked up Persian influences: the prime symbol of their kingship, for example, the crown, was a Persian import, as is the loan word for it in Arabic, taj" (Mackintosh-Smith 2019). Indeed, analyzing the etymology of certain types of words has

long allowed researchers to make inferences about the introduction of certain ideas, technologies, institutions, beliefs, or cultural practices to a particular society. In the quote, for instance, the presence of the new loanword, crown, indicates the source from which new ideas related to kingship have been introduced. However, it is important to note that while it is uncontroversial to interpret the presence of loanwords as—for example—evidence that the Lakhmid concept of kingship was influenced by Persian societies, this does not necessarily mean that they had no prior concept of kingship. Instead, it simply suggests that something new related to this concept has been introduced.

This example relies on the field of etymology, which traces the history of words. Linguists define loanwords as words that have been adopted from another language group, unlike neologisms, which are invented within a given language,[8] and cognates, which are inherited from an ancestral language. Cognates have been most heavily studied by linguists; in particular, the field of glottochronology—where differences in cognates are used to date the age of branches in linguistic family trees (Vansina 1990)—has received considerable attention.

This study of language ancestry is complicated by the possibility of horizontal transmission, which has led linguists in the field of glottochronology to try to exclude loanwords as much as possible. To do so, they compile lists of core meanings that are essentially required in all languages. These words are considered unlikely to have been borrowed, since each language would have very likely had to include some version of them prior to borrowing from another group. These Swadesh lists (first developed by Morris Swadesh) are used in many applications to identify distance between language groups (Swadesh 1950).[9]

While glottochronology seeks to exclude horizontal language transmission, a literature on "wave-like" language evolution (originally proposed in Schmidt (1872), cited in List (2014)) stresses that horizontal transmission is a pervasive source of linguistic differences. The exclusion of horizontal borrowing has been identified as a major limitation of glottochronology, with its strictly "tree-like" models of language evolution. This critique has led researchers to consider new types of data to allow

[8] Linguists use the term loanwords and refer to words as borrowed or loaned, even though they recognize that the lending metaphor is a poor one (e.g., words are non-rival and will obviously not be "returned"). They do this because the terms have come to mean something very specific within the field. In fact, paradoxically, the persistence of this jargon has been attributed to the metaphor being terrible. Since nobody outside of linguistics would naturally refer to words in this way, the formal definitions have not been diluted or corrupted by laymen. We will interchangeably refer to loanwords as being adopted or borrowed.

[9] One such prominent application is the Automated Similarity Judgment Program (ASJP) (Wichmann, Holman, and Brown 2016).

for more complex models that incorporate cross-societal influences (Ben Hamed 2014). Within this literature, historians and linguists regularly interpret the presence of loanwords as evidence of influence.

One notable example of this allows historians to trace cross-societal contact between East and West, dating as far back as the Parthian Empire (circa 247 BCE–224 CE), from an era in which written records are quite difficult to come by. That work concludes that "Buddhism made sizeable inroads along the principal trading arteries to the west […] The rash of Buddhist loan words in Parthian also bears witness to the intensification of the exchange of ideas in this period" (Frankopan 2016, p. 32).

For economists, being able to directly measure the external influences on economic markets or formal institutions could represent an important opportunity to better understand how they evolve. One clear application of this is the work in economics on the impact of colonialism, and there are parallels in linguistics as well. Consider, for instance, the following quote about Swahili, a commonly spoken language across British-colonized East Africa: "English influence is concentrated on the semantic field Modern world, including (modern) clothing and the (modern) legal system" (Schadeberg 2009, p. 87). While economists tend to exploit natural historical experiments to better understand the impact of colonialism, linguists are able to tackle the question more directly by assessing the types of words that were borrowed from colonists. Through this complementary approach, they have been able to identify specific institutions and technologies that were particularly heavily influenced by colonists.

Religious Origins

To validate our empirical methodology requires an idea of the "true" origin against which to compare, but it is worth keeping in mind that the notion of a single "true" religious origin is already an oversimplification in many cases. This issue is further complicated by the fact that the global origin of a religion depends on whether we are considering largely localized oral spread through preaching by sacred figures or global spread, which predominantly took place via the creation of a canonical scripture.

In some cases, these locations are the same, or at least very similar (see Online Appendix A for more detailed accounts of the various modes of early religious spread for the religions we consider). In the case of Buddhism, the preaching of the Buddha, Siddhattha Gotama, was centered in the Ganges River basin near his birthplace in Lumbini (near what is now the Nepal-India border). This region is marked in pink in Figure 1, and the centroid of that region is listed in Table 1, Columns

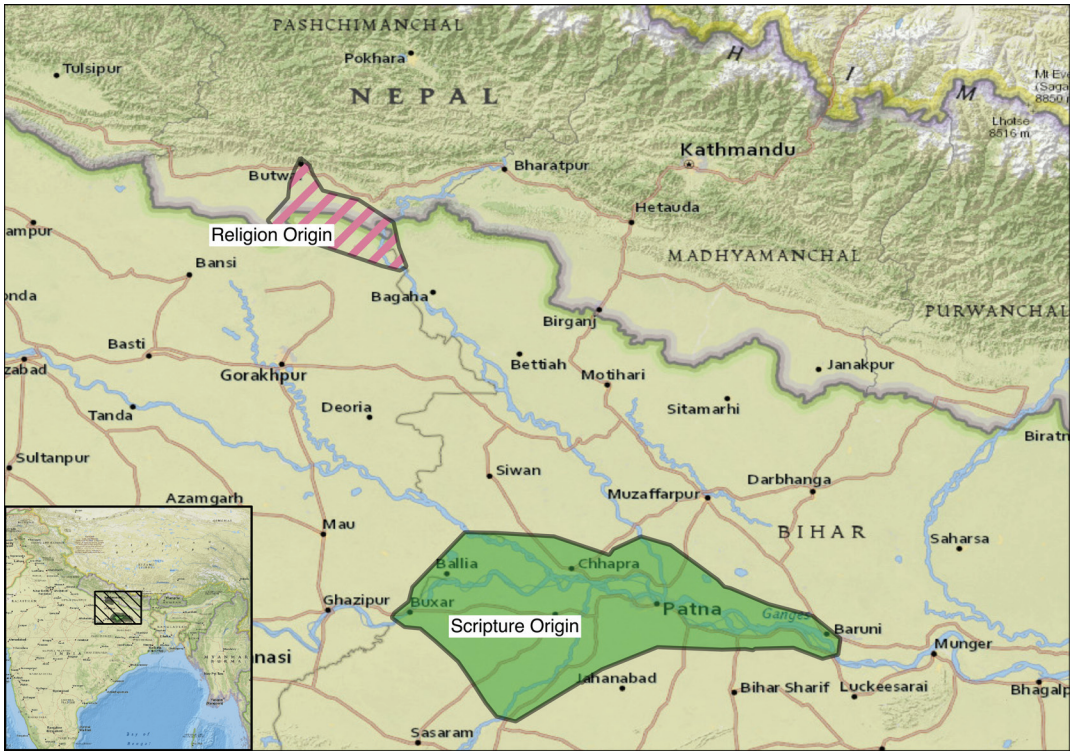


FIGURE 1
MAP OF THE ORIGINS OF BUDDHISM (PINK) AND ITS SCRIPTURE (GREEN)

Note: This map shows the historical account of the geographic origin of Buddhism itself, in pink, as well as the geographic origin of Buddhist scripture, denoted in green. The centroids of these regions are listed in Table 1. See the published online version of this paper for colored figures.
Sources: Authors' generated map.

(3) and (4). The councils of disciples that decided the core scriptures of Buddhism were held near Rājagaha, also within the Ganges Valley. This region is depicted in green in Figure 1, with centroid in Columns (1) and (2) of Table 1. While these regions are not identical, they are very near one another and are close enough that we will have no chance to empirically distinguish between them.

Likewise, the early spread of Islam—both in terms of the preaching of Muhammad and the compilation of early manuscripts of the Qur'an—emanated from a similar place and time. Muhammad was based in Mecca and later Medina in the seventh century CE. We demarcate the historical Mecca and Medina provinces with pink diagonal lines in Figure 2.[10] The Qu'ran, meanwhile, was collected into one volume after the death of Muhammad, by the first caliph, Abu Bakr (r. 632–634). By this time, the Rashidun caliphate comprised the majority of the Arabian Peninsula, and its capital had moved just east of Mecca and Medina, toward contemporary Riyadh (Campo 2009). This is depicted in green in Figure 2, and the eastward movement in the centroid is reflected in Table 1. However, as with Buddhism, the origin of religious spread is not markedly different if we consider where Muhammad was based or where the first scripture was compiled.

The same is not true of either Hinduism, Judaism, or Christianity. In the case of Judaism, the establishment of the Kingdom of Israel and the confederation of the 12 tribes of Judaism occurred in the area west of the Jordan River near Jerusalem (denoted in pink in Figure 3). However, historians believe that canonical Jewish scripture was compiled during exile in Babylon to codify and preserve Jewish religious life and laws (denoted in green in Figure 3). In this case, the origins of religious spread through preaching and the origin of scripture would not be similar. We can see this in Table 1. Column (5) shows that while the origins of scripture and the religion itself are less than 500km away for each of Buddhism and Islam, they are over 1,000km away for each of the other three major religions.

For Christianity, the origin of religious spread via preaching would have been centered on the events in the life of Jesus Christ in and around Jerusalem (denoted in pink in Figure 4). The creation of codified Christian scripture, however, was not centered in the same region as the events depicted in the Bible. Instead, this was driven by later Greek-speaking early Christians, namely Paul, a Greek speaker from modern-day Turkey. Early Christian gospels were also written in Greek, not the Aramaic that would have been spoken by the original disciples. The Bible, meanwhile, was first compiled by in Alexandria, so the spread of scripture would have

[10] These regions are based on the maps in Armstrong (2001).

FIGURE 2
MAP OF THE ORIGINS OF ISLAM (PINK LINES) AND ITS SCRIPTURE (GREEN)

Note: This map shows the historical account of the geographic origin of Islam itself, in pink, as well as the geographic origin of Islamic scripture, denoted in green. The centroids of these regions are listed in Table 1.
Source: Authors' generated map.

FIGURE 3
MAP OF THE ORIGINS OF JUDAISM (PINK) AND ITS SCRIPTURE (GREEN)

Note: This map shows the historical account of the geographic origin of Judaism itself, in pink, as well as the geographic origin of Judaic scripture, denoted in green. The centroids of these regions are listed in Table 1.
Source: Authors' generated map.

FIGURE 4
MAP OF THE ORIGINS OF CHRISTIANITY (PINK) AND ITS SCRIPTURE (GREEN)

Note: This map shows the historical account of the geographic origin of Christianity itself, in pink, as well as the geographic origin of Christian scripture, denoted in green. The centroids of these regions are listed in Table 1.
Source: Authors' generated map.

emanated from the historically Greek regions depicted in green in Figure 4, well west of Jerusalem. Similar to Judaism, the origins of Christian preaching and the origins of Christian scripture are quite distinct.

The nature of the oral and written origins of Hinduism is less clear than those of the other religions we consider, which is unsurprising given it is, by far, the oldest. There is continuing debate on the origins of Hinduism that relates to the uncertainty about the origins of the Indo-European languages. While this is an incredibly complex issue, it is notable for our purposes that, given the age of Hinduism itself, its actual origins are tied to early Indo-European settlements. According to the predominant "steppe hypothesis," this traces back to Early Bronze Age migrants from the Pontic-Caspian steppe, north of modern-day Turkey. Accordingly, we denote this as the religious origin, denoted in pink in Figure 5, and the centroid of that region is used for distance calculations throughout, as reported in Table 1. In terms of the origins of Hinduism's scripture, there are, broadly, two mainstream hypotheses. The first is that it originated in the Bactria–Margiana Archaeological Complex in present-day Afghanistan (the northernmost region denoted in green in Figure 5) and occurred before proto-Indo-Europeans spread south to the Indus Valley. The second hypothesis is that Hindu scripture originates in the Indus Valley and was adopted by proto-Indo-Europeans after they had migrated to this region (the more southern region denoted in pink in Figure 5). There is also a separate hypothesis that Hinduism originated within India; however, this has far less support among historians and is outside of the mainstream view of scholars.

We take the centroids of each of the scripture and preaching origins for each of the five religions we consider and present them, along with the distance between these centroids, in Table 1.

DATA

The foundation of our approach is to quantify and analyze the intensity and direction of religious language transfer among language groups.[11] To accomplish this, we build a dataset on religious loanwords, which requires first identifying a set of religious words, and then assessing which ones were "borrowed" and from whom.

To do this, we start by identifying words related to religion using a list of seed words based on a standard topic classification scheme. Next, we

[11] Here, we use the Ethnologue for our definition of language groups. The study region itself is in Online Appendix Figure C1, while the boundaries of the groups within this region are shown on the map in Online Appendix Figure C2.

FIGURE 5
MAP OF THE ORIGINS OF HINDUISM (PINK) AND ITS SCRIPTURE (GREEN)

Notes: This map shows the historical account of the geographic origin of Hinduism itself, in pink, as well as the geographic origin of Hindu scripture, denoted in green. The centroids of these regions are listed in Table 1.
Source: Authors' calculations.

estimate which words were borrowed from other languages and identify the most likely source language. Finally, we aggregate this word-pair-level data to the language-pair level. This process is based on the methodology described in Blouin and Dyer (2021). We will outline how we identify religious words, and then describe the algorithm for identifying loanwords among these religious words.

Identifying Religious Words

To identify language transfer related to religion, we first need to identify a set of words broadly related to religion. This is a multi-step process, whereby we first identify a set of seed-words in English, then expand this set of seed words to capture all semantically similar concepts in all other languages, and then codify this set of concepts to estimate loanwords. We will describe each of these steps in turn.

SEED-WORDS IN ENGLISH

The task of identifying religious words begins with a small number of seed words in English. We identified seed words by starting from the Library of Congress Classification (LCC) system as an external, objective guide of words and concepts that represent the topic of religion. These words represent the concepts, people, and places of worship in the major religions we aim to represent. They were deliberately selected to cover religious concepts, without prioritizing the means of religious spread or specifically including religious texts.

Our primary motivations for using the LCC were to tie our hands and to be as transparent as possible. An alternative option would have been to compile our own list of seed words tailored to the context, but this would leave a large degree of methodological freedom to search through plausible lists until the desired result is obtained. The LCC is a reasonably objective and widely known classification system, with a relatively complete, neutral, and objective set of classification categories.

We started from the LCC Subclass BL (Codes BL1-2790, Religions, Mythology, and Rationalism), summarized in Online Appendix Table B1.[12] We then removed headings related to Mythology and Rationalism, as well as those related to the study or classification of religions. We also removed headings related to the history of specific religions and specific

[12] Original classification schema sourced from https://www.loc.gov/aba/cataloging/classification/lcco/lcco b.pdf.

TABLE 1
RELIGIOUS ORIGINS

	Coordinates of Possible Origin of Religious Spread				
	Origin of Scripture (Centroid)		Origin of Religion (Centroid)		Scripture - Religion Difference
	Latitude (1)	Longitude (2)	Latitude (3)	Longitude (4)	Distance (km) (5)
Buddhism	84.80	25.57	83.63	27.43	237.34
Islam	46.15	23.18	43.34	21.54	338.22
Hinduism	69.82	34.34	42.79	48.51	3,111.85
Judaism	44.35	32.95	34.84	31.60	1,063.72
Christianity	25.44	35.81	35.18	31.79	1,149.41

Note: This table presents the centroids of the possible origins of religious spread presented in Figures 1–5.
Source: Authors' calculations.

religious doctrines. We dropped any references to technical classification words such as *General*, as shown in Online Appendix Table B2. What was left over after these removals was used as our list of seed words. We cleaned this data by replacing some of the more esoteric terms with synonyms more likely to be found in common language or ones less likely to have non-religious connotations. In both cases, this was done to facilitate the expansion of the seed words in the next step of the process.[13] The resulting seed words, as well as the justifications for any such data cleaning, are in Table 2.

EXPANDING TO OTHER LANGUAGES

With these English seed words in hand, the next priority was to propagate this list across the languages in our sample. We used the English seed-words to identify related words in nearly three hundred languages from around the world, based on semantic similarity. For an overview of this process, the entire routine is presented graphically in Section B.1.1 and Figure B2 of the Online Appendix. The intuition behind this procedure is to look for similar sentence structures across languages to see which words in these other languages are often used.[14] Doing so allows us to mitigate any bias introduced by the English seed-words.

The goal is to propagate the initial list of the seed words across each language group. The data source for language groups throughout is the

[13] Since the seed-word expansion searches Wikipedia for synonyms, it is important that (a) our seed words are common enough to appear on Wikipedia, and (b) are unambiguously religious.

[14] For instance, for places of worship, we might find "Temple" in some languages or "Mosque" in others, which are not direct translations of one another.

TABLE 2
CHOICE OF RELIGIOUS SEED WORDS

Heading	Seed-Words	Justification
Religion	*religion*	This is straightforward word to include, as the word religion is commonly used.
Sacred books	*sacred*	Here we drop the word "book" and keep "sacred," as we do not want to bias toward identifying the spread of books and scripture.
Natural theology	*god, astrology*	The sub-headings for theology focus primarily on deities, and different types of understanding of deities, so "god" is a fairly broad representation of this concept that appears in common usage. We also include "astrology" to capture a broader range of natural theology.
The soul	*spirit*	Here, soul is a commonly used word that is broadly applicable across all of our religions of interest. We selected *spirit* as a seed-word for the soul category, as the concept of a soul is less universal than the concept of a *spirit*.
Eschatology	*afterlife*	Eschatology is defined in the *Oxford English Dictionary* as "The department of theological science concerned with The four last things: death, judgement, heaven, and hell'." In order to represent this without specifically referencing Christian or another specific understanding, we chose to include *afterlife* as a broad seed-word capturing concerns about what happens after death or the ending of the world.
Worship. Cultus	*worship*	As the word "cult" may have other non-religious connotations and may be more likely used in the study of a certain religion rather than by its practitioners, for this category, we chose the word *worship*, which occurs in common usage and is fairly universal.
Religious life	*pray*	For religious life, we chose to include the seed-word *pray*, as the concept and act of prayer appear to be relatively universal across most religions, without including non-religious concepts such as "contemplation" or "meditation."
Religious organization (people)	*priest*	We include the word priest, as well as similar words *monk* and *preacher*, to capture a broad range of people involved in religious organizations.
Religious organization (places of worship)	*church, temple, mosque*	We include these seed words for different forms of religious institutions, including other similar words such as *synagogue*, *shrine*, and *sanctuary* to broadly cover the concept of places of worship.

Sources: Word headings sourced from the Library of Congress Classification. Seed words selected by authors. This table describes how we go from the final list of relevant headings from the Library of Congress Classification in Online Appendix Table B2 to the actual seed words we use for our semantic similarity routine to identify related words across languages.

well-known Ethnologue (Lewis 2009).[15] To expand our seed-words to each of these language groups, we start from data on the words that exist in each language—the lexicon of the language. These lexicons come from PanLex, a single coherent lexical database built from thousands of translation dictionaries and including over 25 million words.[16] PanLex includes most living languages and can be directly matched to the ISO 639-3 codes used in the Ethnologue. These combined word lists include as close as possible to all known words in all known languages. PanLex includes meaning IDs for each word, so as a first step, we can match our English seed words to translations in each other's languages using the meaning identifier. Each of these words is converted into the International Phonetic Alphabet (IPA) using data from Ager (2019) and Mortensen, Dalmia, and Littell (2018), so we can compare words across different scripts.[17]

However, if we stopped at direct translations, we would risk the list of religious words capturing a large Western bias. So, it was important to identify religious concepts in each of these languages, as they are typically used in those languages, rather than being restricted only to direct translations of the English seed words. To do this, we implemented a well-established semantic analysis routine trained on Wikipedia data (see Bojanowski et al. 2017) for 294 languages.

The logic is, for each language, to represent words numerically in a way that captures the meanings of words and how they are associated with each other. The similarity in the contexts in which words are used allows us to compute the "distance" between two words.[18] To do this, we represent words as vector values in a 300-dimensional vector space, where each of these dimensions is intuitively related to a "feature" that captures the relationship between two words. For example, the word "Queen" can be represented as being quite similar to the representation "King - Man + Woman" (Mikolov, Yih, and Zweig 2013).

[15] The goal is to identify religious words in all languages. Throughout the study, a language group, as defined by the digitized Ethnologue map of ethnolinguistic societies, is the unit of observation. The Ethnologue provides the locations of each language, and it includes both contemporary languages as well as recently extinct and vulnerable languages. In the Ethnologue, borders for each group are provided, which allows us to compute the centroid of each group.

[16] PanLex is a non-profit with the mission of improving resources available to underserved languages. To do this, they have attempted to build the largest possible lexical translation database. See https://panlex.org. The database is constantly being updated to include new sources, and for our analysis, we the dataset as it was on 1 October 2018.

[17] For further information on how we filtered out phrases and expressions that are not words, see Section B.1.

[18] This has been used in economics as a way to measure worldviews and cultural discourse (Giorcelli, Lacetera, and Marinoni 2022).

After finding direct translations of the seed words (i.e., those assigned identical meaning identifiers in PanLex) among the covered languages in PanLex, we use this routine to identify words that are not direct translations but are similar. To consider a broad range of associations, we consider two meanings as similar if their word-vector representations are similar to a seed word or its direct translation in any of the languages covered. This means that even if a concept is not closely related to religion in English, but is semantically similar in another language, we can include this association in our list identifying religious words. Therefore, the concepts we identify as related to the initial seed words are not purely based on English worldviews. We take these "similar meanings" and again translate the expanded word set using the PanLex meaning IDs to get a large list of words in each language that are related to religious seed words.[19]

There are several important advantages to this method. The first is that it allows for broader coverage. Some of the languages in PanLex have more coverage than others, and expanding the set of words that we examine increases the odds that one or more of them is included in the less heavily documented languages. Second, it is important not to narrow in too closely on the loanwords data. Our intention was to develop a way to examine *global* patterns in language transmission. Rather than getting into the process of defending the loanword status of specific word pairs—which is the focus of linguists[20]—our approach is to acknowledge that any automated approach will come with errors, and we should accordingly manage those errors to the best of our ability. One way to do this is by exploring averages of larger subsamples, whenever possible. Finally, the procedure aims to minimize the likelihood that—despite the relatively objective nature of the LCC—our identification of religious words is driven by word associations in English and hence reflects solely Western worldviews.

Once we have identified all similar words in all languages in the Ethnologue, both the original English seed-words and the much larger set of semantically similar words in the other languages are matched to the meaning IDs described earlier. This comprises our final list of religious words.

[19] This produces a list of over 8,000 meanings that are associated with our original English seed words. The vast majority do not have direct English equivalents, but we present in Online Appendix Table B3 the English words associated with these additional meanings.

[20] We view our approach as complementary to the work that linguists do. It is certainly not a substitute, since we cannot claim with anywhere near the same level of certainty that any particular word pair is, or is not, a loanword pair.

Machine Learning Algorithm: Identifying Loanwords

Having algorithmically identified a set of religious words across the world's languages, the next step is to identify which of these words were borrowed from other languages, and to identify the source language. While Panlex is a near-complete list of words in the world's languages, it does not contain the necessary information on borrowing. To generate this data, we use a standard machine learning algorithm to predict loanword status and identify the most likely source. Our approach was to automate the procedure used by linguists to identify loanwords as closely as possible. To this end, we follow the discussion of this process in the section *Recognizing Loanwords* from the authoritative guidebook *Loanwords in the World's Languages: A Comparative Handbook* (Haspelmath and Tadmor 2009). To the extent that is possible, we aimed to create computational analogues based on Haspelmath and Tadmor (2009) to generate features in our data set that approximate the features that linguists typically consider.

However, to do this, we needed a validated set of loanwords we could use to train the classifier. This data does exist, in the form of the World Loanword Database (WoLD), which is the largest dataset of consistently compiled loanwords identified by linguistic experts. To be more precise, WoLD includes "vocabularies (mini-dictionaries of about 1,000–2,000 entries) of 41 languages from around the world, with comprehensive information about the loanword status of each word," and identifies the source words for these borrowings from 369 other languages. We used this data set to train our machine learning algorithm on the word-pairs in PanLex that can be matched to WoLD. We then applied the classifier to all of the word-pairs in PanLex that are potential loanwords.

To do this, we started by creating a word-pair level database of words that are semantically similar and thus may have been transferred from one language to another. An overview of the process, along with the databases and tools used at each stage, is presented in Online Appendix Figure B1. To build the training set, we drew a stratified sample from the subset of PanLex word-pairs that are also included in WoLD.[21] We had to address the fact that the training set is heavily imbalanced, with many fewer true loanword word-pairs than non-loanword word-pairs. This poses a problem because it could result in high accuracy by drastically underestimating loanwords. We dealt with this by selecting only a

[21] This stratified sample included some word-pairs that were actual loanwords, and different types of non-loanword word pairs including non-borrowed words, borrowed words matched to the wrong source word, and borrowed words where the direction of borrowing is inverted.

random subsample of the heavily overrepresented categories and then augmenting the underrepresented categories with synthetic oversampling (Chawla et al. 2002; Lemaitre, Nogueira, and Aridas 2017).[22] Based on this training set, we predicted loanword status using a random forest classifier. Estimation details are in Appendix Section B.1.2. Overall, the accuracy of the classifier was approximately 98 percent.[23]

After training the classifier, we applied it to the full set of potential loanword word-pairs in PanLex, selecting the highest-probability source word for each.[24] We then restricted to the set of words identified as religious words (as described in Section B.1.1 in the Online Appendix) and constructed measures of intensity of religious borrowing between language pairs. This aggregated variable represents language adoption by group i from group j and is defined as follows:

$$L_{ij} = \frac{\#ReligiousLoanword_{ij}}{\#ReligiousWord_{i}} \tag{1}$$

We define $\#ReligiousWord_{i}$ as the number of religious words in the language of society i. Similarly, $\#ReligiousLoanword_{ij}$ is the number of religious loanwords in the language of society i originating from j. L_{ij} is therefore the share of religious words in society i that were adopted from society j, or equivalently, a measure of the religious linguistic influence of j over i. It is worth noting that L_{ji} is a separate observation indicating religious linguistic influence in the opposite direction, of group i over j.

Summary statistics are in Table 3. They show that, conditional on any language adoption, borrowing between a typical language-pair accounts for approximately 3 percent of religious words. We use this pairwise data to construct a directed network of religious language transfer among Ethnologue groups. Details of how we construct the networks and associated measures of network centrality are in Online Appendix B Section 2.

[22] We implemented the classification procedure in two stages, with a coarse first-pass to remove obvious non-loanwords, and a second-stage refined classifier that focused on the less-obvious cases, such as cognates vs. loanwords or loanwords with the direction of transfer being inverted. We then applied this classifier to a much larger subsample and trained a second, more refined classifier on those identified as plausible potential loanwords by the first classifier.

[23] The vast majority of potential loanword word-pairs were rejected by the first-stage coarse classifier. The refined classifier was approximately 92 percent accurate on the less-obvious cases that were not rejected in the first pass and made it to the refined second-stage classifier. We present further details on classifier performance in the confusion matrix in Online Appendix Figure B3.

[24] Please see Online Appendix B.1 for further details on the classification procedure and the features used at each stage.

TABLE 3
SUMMARY STATISTICS

Variable	Mean	Std. Dev.	Min.	Max.	N
Any religious language adoption	0.016	0.124	0	1	4,839,955
Share adopted (conditional on any adoption)	0.03	0.056	0	1	12,910
Distance between lender and borrower centroids (km)	8.206	4.556	0	20.029	4,839,955
Centrality of lender in religious language network	0.005	0.021	0	0.309	4,839,955
Centrality of borrower in religious language network	0.001	0.011	0	0.309	4,839,955
Number of religious words identified	69.543	106.578	0	3438	4,839,955
Latitude of centroid of lender	18.306	13.569	0.078	59.941	4,839,955
Longitude of centroid of lender	50.912	34.525	10.017	109.985	4,839,955
Latitude of centroid of borrower	6.641	18.324	–51.635	73.135	4,839,955
Longitude of centroid of borrower	52.666	83.585	–173.925	177.657	4,839,955

Note: In this table, we present summary stats of the pairwise religious language adoption used to reconstruct our estimates of religious origins. This includes summary statistics for the level of pairwise religious adoption, as well as the network centrality measures of borrower and lender nodes and their coordinates of group centroids. We also share summary statistics of the number of words identified as being religious by the semantic similarity routine, with histograms of the relevant distributions presented in Online Appendix Figures C4 and C5.
Sources: Data on religious words, language borrowing, and religious language networks constructed by authors. Data on distances and coordinates constructed from the Ethnologue.

EMPIRICAL METHODOLOGY

Our empirical approach is inspired by Barjamovic et al. (2019), who collect exceptionally rich historical data on inter-city trade flows to reconstruct the probable locations of "lost" ancient cities. In many cases, collecting such data is not feasible or even possible. One insight of this article is to show that data on language can help with geolocation as well, albeit for slightly different purposes.[25] However, accommodating this broader range of settings introduces various challenges that require non-trivial adaptations of the methodology, so the two approaches should be considered complementary.[26]

Calibration

The empirical exercise begins with the constructed measure of influence (or centrality) in the network of adoption of religious words based on the loanword data described earlier.[27] Using this measure, we estimate

[25] Data and code replication package is available online at https://doi.org/10.3886/E233003V1 (Dyer and Blouin (2025).

[26] Barjamovic et al. (2019) estimate a gravity trade model with commercial records from 12,000 clay-tablets dating back to the nineteenth century BCE, which required an understanding of an Old Assyrian dialect of ancient Akkadian. Without this information, we rely on unsupervised machine learning to separate estimated source points into clusters corresponding to specific religions.

[27] Again, see Online Appendix B for details of construction.

the relationship between religious language influence and distance to a known origin of spread. We then use this information to make inferences about the geographic locations of unknown origins of spread.

For the purpose of validation, this means that we would like to understand, for each of Islam and Buddhism—the two religions with clear and uncontested origins—if we can use what we know about one to estimate the location of the other. To better understand what the methodology is capturing, we use calibrations from both Buddhism and Islam to estimate whether the resulting estimates for each of Christianity, Judaism, and Hinduism are nearer to the origins of the religions themselves or to the origins of the scripture. Across both exercises, the results using either Buddhism or Islam to calibrate are not materially different.

Starting with the validation exercise, we first calibrate using Islam and use this information to estimate the location of Buddhism. Then, we calibrate using Buddhism and estimate the location of Islam. To generate the estimates used for calibration, we proceeded with the regression model in Equation (2). Throughout this paper, we refer to language influencers—the group that is the source of loanwords—and language adopters, the group that adopts the loanword from another language. As before, we denote this using subscript *i* to indicate a language in its role as an adopter and *j* to indicate a language as an influencer.[28]

$$log(d_j) = \beta \mathbf{c} + \gamma\, LexiconSize_i + f(DistanceBetweenGroups_{ij}) + \varepsilon_{ij} \quad (2)$$

In Equation (2), **c** is a matrix containing some polynomial of c_j, which is a measure of linguistic influence. We consider a cubic specification in the main results, but all results are consistent using linear and quadratic specifications as well, and estimates from these models are presented in Online Appendix C throughout.[29] c_j measures influence within a directed network of religious word spread. For the main results, we use eigenvector centrality, which is defined formally in Equation (9) in Appendix B.2. Again, though, results are robust to using alternate measures of

[28] Given that our data is at the directional pair level, each language will appear both as lender and borrower.

[29] c_j is included as a cubic polynomial in the main specification in order to account for the expected pattern of non-linearities in the relationship between distance, lending, and borrowing. For instance, we expect that very nearby an origin is likely to almost exclusively lend and therefore have high out-group centrality, but we expected that this may likely trail off quickly, and those beyond even relatively small radii from the origins (relative to the study region) may almost exclusively borrow. Beyond this, borrowing too would dwindle as religious influence decreases with distance to the given origin. We also wanted to keep the specification consistent for both borrowers and lenders, and felt that including a more flexible specification would make that more sensible.

network influence, which are also presented in the appendix throughout. $f(DistanceBetweenGroups_{ij})$ is the distance between the influencing and adopting language groups. We control for the size of the lexicon included in the source data ($LexiconSize_i$) to account for the possibility that centrality is artificially low when data is sparser.[30] $log(d_j)$ is the natural logarithm of the distance from the centroid of language group j to either Mecca or Lumbini, and results are all robust to modeling this linearly as well.

We do the same for adopters in the network (i.e., those being influenced). In this case, we have a regression equation as follows:

$$log(d_i) = \beta \mathbf{c} + \gamma\, LexiconSize_j + f(DistanceBetweenGroups_{ij}) + \varepsilon_{ij} \quad (3)$$

Everything is defined as before, but the subscripts are swapped. In this case, because the focus is on adopters, the matrix **c** contains elements c_i to measure a language group's propensity for adoption within the network of religious word spread. An observation is a language pair ij. Of course, for all i or j in these regressions, both the network centrality and the distance to Mecca / Lumbini only vary at the group-level, and not the group-pair level.[31] This has implications for the standard errors, so to account for this, they are two-way clustered by groups i and j.

The resulting estimates are in Table 4. We show estimates using Buddhism in Columns (1) and (2) and using Islam in Columns (3) and (4). Importantly, across all specifications, we see significant non-linearities, which partly justify the non-linear specifications in Equations (2) and (3). Again, though, estimates are robust to alternative specifications as well.[32]

Overall, as one might expect, those who are influential within the religious network for Buddhism are nearer to the origins. This can be seen in

[30] As described in Online Appendix B Section "Language Data," our borrowing/lending data is based on the wordlists in the PanLex lexicon for each language, from which we calculate $LexiconSize_i$ (the number of single-word expressions) to control for data availability. We discuss the potential bias from the sources used to construct our data in B.6.

[31] Another valid option would have been to aggregate the data to the group level prior to running the regressions instead of after. The two options are essentially equivalent. However, the next step of converting the predicted distances from these regressions to origin coordinates necessarily takes place at the pair level. Therefore, in this case, we would have to aggregate the data for this step, dis-aggregate for the next step, and then re-aggregate again after that, which seemed unnecessarily complicated. However, the clear trade-off is that in this case, we have a group-pair data set with primarily group level variation. There are the same number of observations for each group in our "stacked" data-structure (i.e., all observations are equally weighted regardless), so the only implication is for the standard errors.

[32] We show plots of actual and estimated distances to Mecca for lenders and borrowers in Online Appendix Figure C5 and show the relationship exhibits the expected pattern.

TABLE 4
CALIBRATION: LINGUISTIC NETWORK INFLUENCE IDENTIFIES GEOGRAPHIC ORIGINS OF SPREAD

	log(Distance to Lumbini)		log(Distance to Mecca)	
Dependent Variable:	Influencer (1)	Adopter (2)	Influencer (3)	Adopter (4)
Network influence - religious words	34.43***	– 3.59**	–16.79***	–16.54***
	(2.10)	(1.47)	(1.87)	(2.57)
(Network influence - religious words)^2	–371.98**	–19.88**	148.37***	70.38***
	(39.85)	(8.75)	(25.58)	(16.11)
(Network influence - religious words)^3	1007.07***	25.66**	–81.25 ***	–647.3***
	(158.08)	(11.58)	(86.04)	(22.22)
Number of Words	✓	✓	✓	✓
Distance between partners (cubic)	✓	✓	✓	✓
N	4,839,955	4,839,955	4,839,955	4,839,955
R^2	0.158	0.184	0.107	0.4793

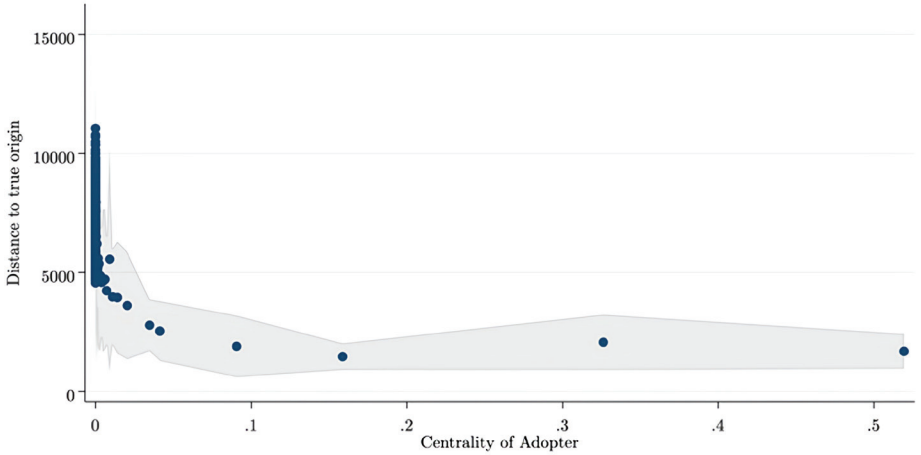
* = Significant at the 10 percent level.
** = Significant at the 5 percent level.
*** = Significant at the 1 percent level.
Note: This table examines the relationship between network influence for religious words and the distance to the origins of religious spread. The unit of observation is a language-group pair. Standard errors are two-way clustered by each language group in the pair.
Source: Authors' calculations.

Figure 6, which presents the scatterplot between network centrality and distance to origin for both lenders and borrowers. The graphs for each show the heavily non-linear relationship that implies that by far most linguistic exchange takes place near the religious origins, either the scripture or the religion.

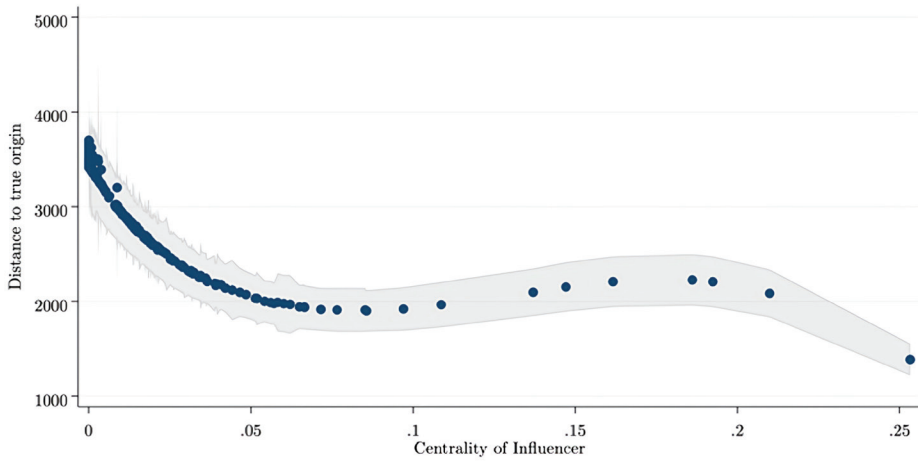
Solving for the Origins of Religious Spread: Euclidean Formula

The next step to solving for the origins of religious spread is to use the estimates from Equations (2) and (3)—which are shown in Table 4—to compute the predicted distance to the origin for each observation. Intuitively, this represents a weighted average of religions, so that, for example, heavy Buddhist influence "pulls" the predicted origin to the east, and heavy Islamic influence "pulls" it to the west. Each predicted distance value minimizes the error from Equations (2) and (3).

It is simple to compute these predicted distances for each influencer and adopter in the data (i.e., using Columns (1) and (2) of Table 4, in the case of Buddhism); however, what we are interested in is geographic coordinates, not distances. These origin coordinates are relatively straightforward to derive from the distances. For each language-pair in the data,



(a) Relationship between network centrality and distance for borrower



(b) Relationship between network centrality and distance for lender

FIGURE 6
SCATTERPLOT OF RELATIONSHIP BETWEEN DISTANCE TO ORIGIN AND NETWORK CENTRALITY

Note: The figure displays binned scatterplots to show the relationship between the network centrality measures for each language group and their distance from the religious origin. The plots are constructed based on 1,500 bins in each case.
Source: Authors' calculations.

the distance represents the radius of a circle emanating from their own language group's geographic centroid. Along the circle formed by this radius lies the estimated religious origin centroid described previously. To convert our radii into a latitude and longitude of this origin centroid, we solve for the geographic coordinates that best rationalize the two

circles (i.e., one estimated for group *i* and the other for group *j*, of pair *ij*). Given our distance estimates ($\hat{d}$) from Equations (2) and (3), these radii are already estimated. The associated coordinates are directly implied by the Euclidean distance formulas. These are:

$$\widehat{d_j} = (10000/90)\sqrt{(\phi_j - \phi_o)^2 + \left(cos\left(\frac{37.9\pi}{180}\right)^2 (\lambda_j - \lambda_o)\right)^2} \qquad (4)$$

and

$$\widehat{d_i} = (10000/90)\sqrt{(\phi_i - \phi_o)^2 + \left(cos\left(\frac{37.9\pi}{180}\right)^2 (\lambda_i - \lambda_o)\right)^2} \qquad (5)$$

In these equations, $\widehat{d_j}$ *and* $\widehat{d_i}$ are the predicted distances based on Table 4. ϕ represents longitude, so that ϕ_j is the longitude of the influencing group (which is known from the Ethnologue), and ϕ_i is the longitude of the adopting group (also known from the Ethnologue). ϕ_o is the longitude of the origin, which is what we would like to solve for. Likewise, λ represents latitude for either group *i* or *j* (both known from the Ethnologue), or origin *o* (which we aim to solve for).

Equations (4) and (5) therefore represent a system of two equations and two unknowns. The two unknowns are the latitude and longitude of the origin $\{\phi_o, \lambda_o\}$. The solution would be trivial if the radii intersected at only a single point (i.e., they were always exactly tangential) since there would be a unique analytical solution. But, of course, this is not always the case due to measurement errors in each of $\{\phi_i, \lambda_i\}$, $\{\phi_j, \lambda_j\}$, and c_i and c_j. Accordingly, we solve numerically for the latitude and longitude that best fit this system using the non-linear estimation procedure outlined in Ross (1990).[33]

This provides us with an estimate of the coordinates of the center of religious influence for each language pair. The estimation procedure converts radii into coordinates, but these coordinates have a similar interpretation to the predicted distance measures we described previously. In other words, conceptually, neither $\widehat{d_j}$ and $\widehat{d_i}$, nor the associated implied coordinates, identify any particular religious origin. Instead, they identify a centroid of origins. Intuitively, this means that if a language were equally influenced by both Islam and Buddhism, both $\widehat{d_j}$ and the associated

[33] For computational efficiency, we implemented this with a 10 percent random sample of the data, which took about three days.

$\{\phi_o, \lambda_o\}$ would represent a convex combination of each origin—which may be far away from both. The more that influence or adoption is confined to a single religion, the closer these distances will get to a true religious origin. Even if influence / adoption within a language pair is mostly skewed toward a single religion, we will end up with clusters of coordinates near the religious origins, rather than the goal of a single point-estimate.

To resolve this issue, we aggregate the estimated coordinates using k-means clustering. We use several other aggregation methods as well, and these produce similar results; they are shown in the appendix throughout. We specify that there should be five origins of spread corresponding to the five global religions (details are in Appendix B.3).[34] Online Appendix Figure C6 shows the efficacy of the k-means clustering routine when we specify a number of clusters different from five. That analysis suggests that specifying five clusters performs best, as it features the lowest rates of mis-assignment of observations to clusters.[35] This implies that even if we had not known to look for five religious origins, and instead used an algorithm to search for the optimal number of clusters, we would have arrived at the same set of five estimates. In addition to this, one of the robustness checks we use is to aggregate using Ward clustering, which is computationally demanding, but does not require a prespecified number of clusters. This method also produces five centroids associated with the five major religions. In all cases, the mean coordinates within each cluster produce five sets of latitude-longitude pairs that correspond to the origins of religious spread for each of the five religions we are interested in.

To benchmark these estimates for the purpose of validation—the exercise using Islam and Buddhism—we follow the exact same procedure outlined earlier, but we replace the language network data with a random number on the same scale.[36] This procedure helps to ensure that we do not accidentally induce a mechanical relationship either through the clustering routine or the choice of study region. If the loanwords-based estimates are systematically closer to the historical account than this benchmark, this can be interpreted as evidence that there is historically relevant information encoded within a society's language.

For the empirical test using Christianity, Judaism, and Hinduism, we employ a similar framework; however, we compare the estimated distances to the origins of scripture to the origins of the religion itself.

[34] We also use alternate clustering methods, as described in Section B.4.

[35] An observation is defined as mis-assigned in the conventional way—when the clustering algorithm assigns it to a cluster that it is not nearest to.

[36] The clustering algorithm is restricted to latitudes between 17.5 and 42.5, and longitudes between 20 and 95. This is to avoid the confounding effects of religions for which we are not trying to pinpoint an origin.

These are essentially the same for Islam and Buddhism, so this exercise is not possible for those (see Table 2). Likewise, validation using Christianity, Judaism, and Hinduism is not possible since the origins of spread for those three religions are not straightforward.

We proceed first with the validation exercise, and then move on to trying to understand whether the language-estimates are capturing the origins of religions themselves or the origins of scripture.

VALIDATION: IS THERE INFORMATIONAL CONTENT EMBEDDED IN LANGUAGE?

Empirical Test

We are interested in two main empirical validation exercises. The first is to compare the locations taken from historical accounts of the origins of religious spread to the model estimates for the same locations (and associated confidence regions). For this comparison, if our model is valid, we expect to be unable to reject the null-hypothesis that these locations are the same. The second exercise is to compare the model estimates that rely on language information to the benchmark estimates that do not. In this case, if we can reject the null hypothesis that the distances to each of the historian's accounts of the origins are the same, then we can conclude that there is relevant information contained in language. Both exercises are important for validating the practice of inferring history from etymology.

Our aim is to see if a purely data-driven approach will correctly fail to reject these reasonable hypotheses.[37] Since a core element of our empirical approach is the *failure* to reject the null, we follow Barjamovic et al. (2019) by reporting confidence areas that are much tighter than the standard 95 percent. In this case, we simply follow Barjamovic et al. (2019) and report 75 percent confidence areas, which makes the region much smaller, and therefore makes it more likely that whenever we do fail to reject the null, that we do so because the estimates are indeed quite similar and not due to noisy estimates.

Another implication of the empirical approach is that we must accept some error. There are a few obvious sources of error, and likely more. One example is that the estimated origins are based on language group regions. The measured locations of these language groups are centroids of

[37] Admittedly, what it can reject may often be more interesting, and we will discuss some of this as well. But, again, our main goal is validation, and getting close to well-established hypotheses is arguably the most convincing way to do this.

geographic polygons and not population hubs,[38] so we should expect this to introduce some measurement error. A second example is that we are unable to observe language dynamics over time. Again, each of these sources of error reduces the precision of the results, but would only represent a source of bias if our inability to account for them systematically moved the estimates nearer to the mainstream hypotheses of religious origins in the history literature. It is difficult to see how this would be the case.

Results

For the validation exercise, we focus on estimates of Buddhism and Islam. In our examination of Buddhism, we rely on the calibration exercise using Islam, and vice versa in our examination of Islam. This is to avoid a mechanically precise estimate of a location based on its own calibration. A map of the results can be seen in Online Appendix Figure C7.

The origins of Buddhism and its spread are historically uncontested. It began in Lumbini in Nepal and spread geographically from Rājagaha near the India-Nepal border, where Buddhist scripture was first compiled (Appendix A.1). The map in Online Appendix Figure C7 displays the historiography-based origin of the religion and scripture in pink and green, respectively, and the estimated 75 percent confidence area with a circle—computed as in Barjamovic et al. (2019). For Buddhism, the confidence area completely overlaps with the areas of historical consensus, indicating that the estimated locations are not significantly different from the actual locations.[39,40] In Table 5, we present the estimated distances to the "actual" origins based on the history literature. We see the estimates for Buddhism in Columns (1) and (2). When we estimate the origins of Buddhism by calibrating with Buddhism (Column (1)), we estimate a difference in only 393km; however, this may obviously be a mechanical relationship. Indeed, in Column (2), where we estimate the origins of Buddhism calibrated using Islam, the estimate is further away, but only slightly. In this case, the estimate remains only 405km away from the true origin. This is much closer than the comparable estimate that excludes linguistic information (nearly 1,400km away). Using the more reasonable Islam-based calibration, the

[38] That is, the centroid of the language polygon could be a location where nobody lives.

[39] We present a series of robustness checks in the Online Appendix Figures C8 and C9 as well. C8 is for Buddhism, C9 for Islam. In each, the (a) subfigure shows that the estimated coordinate is in essentially the same place when we calibrate using a linear specification. In (b), we show robustness to a quadratic specification. In (c), we calibrate using a linear dependent variable instead of the log-dependent variable. In (d), we calibrate using Betweenness Centrality instead of Eigenvector Centrality, while in Є we examine Degree Centrality instead of Eigenvector Centrality. In all cases, the estimated origin location is essentially unchanged.

[40] In Online Appendix Figures C10 and C11, we also demonstrate robustness to various alternate clustering algorithms. Again, the estimated origin location is essentially unchanged.

TABLE 5
VALIDATION: IS HISTORICALLY RELEVANT INFORMATION EMBEDDED WITHIN LANGUAGES?

Religious Origin:	Buddhism		Islam	
Calibration Using:	Buddhism (1)	Islam (2)	Buddhism (3)	Islam (4)
Mean distance (km) using religious loanwords-based estimates	393.2	405.8	392.2	287.9
Mean distance (km) using random estimates	1,438.4	1,387.2	867.5	1,491.8
Difference (km): random - loanwords	1,045.2	981.4	475.3	1,203.9
t-statistics - H_0: random - loanwords = 0				
Regular t-statistic	106.5***	232.9***	109.6***	304.8***
N	9,533	18,001	9,456	23,243

Note: In this table, we present the distances between the centroids of the true origins of Islam and Buddhism and the estimated ones. We do this for both the estimates derived from the calibration exercise (based on Table 3) as well as based on random information in place of the calibration. In Columns (1) and (2), we show the estimates for Buddhism, calibrated based on the distance to Buddhism (Column (1)) and the distance to Islam (Column (2)). In Columns (3) and (4), we show the estimates for Islam, calibrated based on the distance to Buddhism (Column (3)) and Islam (Column (4)). Toward the bottom of the table, we compute the difference between the differences based on the linguistic network calibration and the random information estimates, and present t-tests for the null-hypothesis that the estimates based on random information are the same as those based on linguistic network information. In all cases, we can reject the null, on the basis that the distances based on language information are always smaller than those based on random information. The number of observations changes from column to column based on the number of estimates assigned to each respective cluster.
Source: Authors' calculations.

language-based estimate is more than three times closer to the true origin, a difference that is significant well beyond the 1 percent level.

Second is Islam. The origin of the religion itself is the portion of the Arabian Peninsula under the rule of Muhammad at the time of his death, while we take the area under the rule of Abu Bakr as the region of origin of the written scripture. The maps in Online Appendix Figure C7, just as with Buddhism, show an almost complete overlap between our estimated regions and the historiography-based consensus regions. This implies that there is no significant difference between our estimates and the historical account.[41,42] Furthermore, in Columns (3) and (4) of Table 5, we present the precise distances between our estimates and the historical consensus.

[41] We present a series of robustness checks in the Online Appendix Figure C9 as well. In Figure C9a, we show that the estimated coordinate is in essentially the same place when we calibrate using a linear specification. In Figure C9b, we show robustness to a quadratic specification. In Figure C9c, we calibrate using a linear dependent variable instead of the log-dependent variable. In Figure C9d, we calibrate using Betweenness Centrality instead of Eigenvector Centrality, while in Figure C9e, we examine Degree Centrality instead of Eigenvector Centrality. In all cases, the estimated origin location is very near the original estimate and not significantly different from it.

[42] In Online Appendix Figure C11, we also demonstrate robustness to various alternate clustering algorithms. Again, the estimated origin location is essentially unchanged.

These estimates for Islam paint a very similar picture to the estimates for Buddhism. We show the calibration based on Buddhism in Column (3) and the one based on Islam in Column (4). As before, we include both for completeness, but there is something mechanical about estimating Islam's origins using Islam-based calibrations. This is reflected in the estimated distance, just as with Buddhism, so we focus on the larger Column (3) estimate, which presents the estimate of the origin of Islam, calibrated using Buddhism. This estimate is actually very similar to the Buddhism estimates we saw in Columns (1) and (2), and off from the true origin by only 392km. In contrast, the estimate based on an identical procedure with the exception that we omit information on language, leads to an analogous distance of over 850km. The difference between these estimates is significantly different from 0, well beyond the 1 percent level.

Overall, both the Buddhism and Islam estimates are very close to the historical account, and in both cases, the estimates can be statistically assessed as more informative than estimates lacking any linguistic information. This implies that there is historically relevant information embedded in language, and this information can be leveraged to make inferences about history when records are lacking.

APPLICATION: IS GLOBAL SPREAD DRIVEN BY RELIGIOUS FIGURES OR SCRIPTURE?

While there appears to be important information embedded within a society's language, what that information reflects remains unclear. This question is crucial since, while we can accurately estimate religious origins in the two most straightforward cases, Islam and Buddhism, we should still acknowledge that there have historically been divergent conclusions based on linguistic and archaeological evidence. So far, we have no way of providing insight into whether these discrepancies are due to inherent bias in the analysis of linguistic data (Coleman 1988; Diebold 1994; Lehmann 1968) or because the two approaches are inherently measuring the origins of different phenomena. The intuition behind the latter possibility is that the linguistic approach focuses on *spread*, while archaeological evidence identifies the presence of the societies themselves. These may be the same locations, but may not be. It seems possible that any loanword-based approach is more likely to estimate the origin of this spread rather than the origin of the religion itself. In the case of religion, these locations happen to be very nearby in the two cases we have looked at so far, but this is not the case for Judaism, Christianity, or Hinduism. Accordingly, we now apply our approach to the origins of

these three religions, with an eye toward whether they are identifying the origin of scripture or of the religion itself.

Starting with Judaism, it is widely agreed that the religion itself developed in Jerusalem. However, scripture was either conceived and written (in the case of the Talmud) or codified (in the case of the Torah) near Babylon—the capital of ancient Babylonia. Babylon is where the Jewish aristocracy was exiled by Nebuchadnezzar (Appendix A.4) and was a central hub of Jewish life for over 1,000 years since. These locations are denoted in Online Appendix Figure C12, where the region around Babylon is in green, and Jerusalem is in pink.

The distance between our estimated location for Judaism and the locations of both Babylon and Jerusalem can be seen in Table 6, Columns (1) and (2). In Column (1), we present the distances calibrated using Buddhism, and in Column (2), they are based on the Islam estimates. The distances are consistently quite close to each other (within about 200km), which reflects that the coordinates estimated using each are quite similar (Online Appendix Figure C12). In both cases, the estimates are much closer to ancient Babylon than to Jerusalem. When we calibrate with Buddhism, in Column (1), the distance to Babylon is more than 4 times closer to our estimate than the distance to Jerusalem, and more than twice as near when we calibrate using Islam. In both cases, therefore, the estimates favor the origin of the scripture over the origin of the religion itself. The difference between these two distance estimates is statistically significant well beyond the 1 percent level. That said, the difference between the history-literature consensus origin of scripture and our estimate is not significantly different. While our 75 percent confidence areas are much larger than the true regions, the two areas completely overlap with both calibrations (Online Appendix Figure C12). This is not the case for the origin of the religion itself, where there is no overlap at all when we calibrate with Buddhism (Online Appendix Figure 9a), and only partial overlap when we calibrate with Islam (Online Appendix Figure C12b).[43,44]

Next, we turn to Christianity, which presents a similar dilemma to Judaism. Did Christianity primarily spread from Jerusalem, where Jesus

[43] We present a series of robustness checks in the Online Appendix Figure C13 as well. In Figure C13a, we show that the estimated coordinate is in essentially the same place when we calibrate using a linear specification. In Figure C13b, we show robustness to a quadratic specification. In Figure C13c, we calibrate using a linear dependent variable instead of the log-dependent variable. In Figure C13d, we calibrate using Betweenness Centrality instead of Eigenvector Centrality, while in Figure C13e, we examine Degree Centrality instead of Eigenvector Centrality. In all cases, the estimated origin location is essentially unchanged.

[44] In Online Appendix Figure C14, we also demonstrate robustness to various alternate clustering algorithms. The estimated origin location is essentially unchanged across different clustering methods.

lived? From Constantinople (i.e., Istanbul), where Christianity was institutionalized? Or from North Africa, where the New Testament was written and canonized (Appendix A.5)? Christianity is even slightly more difficult to deal with than Judaism because, even if we only consider the origin of scripture, it is not entirely clear what the appropriate origin location should be. For instance, Alexandria appears to be a reasonable choice, as the location where the New Testament was compiled. But equally reasonable could be Greece, where Paul proselytized and wrote the majority of the early chapters of the New Testament. Because of this, in Online Appendix Figure C15, we represent the region surrounding the Mediterranean in green to represent the origin of scripture, while we denote Jerusalem, the origin of the religion, in pink.

Nevertheless, the estimate remains much closer to the monastic centers at the time of Christianity's spread than it does to the religion itself. This can be seen most clearly in Table 6, which shows the distances from our estimate to each of the religious origins and the origin of the scripture (Columns (3) and (4)). When we use the Buddhism calibration (Column (3)), the distance to the scripture's origin is just over half the distance to the religion's origin, whereas when we use the Islam calibration (Column (4)) the distance to the origin of the scripture is about 2.5 times closer. Both of these differences are statistically significant beyond the 1 percent level, as they were in the case of Judaism.[45,46]

Finally, we move to Hinduism. In the case of Hinduism the historical account is far from resolved (Appendix A.2). The ongoing debate attributes the origins of Hindu scripture either to the Indus Valley civilization (in the Indus Valley), where archaeological evidence has found similarities with iconography in modern Hindu scripture, or to central Asia, where the oldest known Hindu scripture, the *Rg veda*, has been attributed. The origin of Hinduism itself, though, is incredibly old, by far the oldest of the five religions. The debate is contentious because it is tied to the origin of Indo-European people, which itself remains a heavily-debated academic question. However, the most dominant hypothesis places the origin in the Pontic Steppe.

[45] We present a series of robustness checks in the Online Appendix Figure C16 as well. In Figure C16a, we show that the estimated coordinate is in essentially the same place when we calibrate using a linear specification. In Figure C16b, we show robustness to a quadratic specification. In Figure C16c, we calibrate using a linear dependent variable instead of the log-dependent variable. In Figure C16d, we calibrate using Betweenness Centrality instead of Eigenvector Centrality, while in Figure C16e, we examine Degree Centrality instead of Eigenvector Centrality. In all cases, the estimated origin location is essentially unchanged.

[46] In Online Appendix Figure C17, we also demonstrate robustness to various alternate clustering algorithms. Again, the estimated origin location is essentially unchanged.

TABLE 6
WHAT DOES LANGUAGE CAPTURE? ORIGIN OF SCRIPTURE OR RELIGION

Religious Origin:	Judaism		Christianity		Hinduism	
Calibration Using:	Buddhism (1)	Islam (2)	Buddhism (3)	Islam (4)	Buddhism (5)	Islam (6)
Mean distance (km) to religion origin	1000.2	1,085.8	822.1	726.9	2,757.0	2,937.2
Mean distance (km) to scripture origin	221.9	433.2	448.2	283.9	470.5	1,132.2
Difference (km): religion - scripture	530.1	778.2	373.9	443.0	1,613.8	1,805.0
t-statistics — H_0: religion – scripture = 0						
t-statistic	34.8***	90.9***	27.4***	57.3***	188.2***	228.8***
N	6,626	20,944	8,224	25,712	10,410	24,684

Note: In this table, we present the estimated latitudes and longitudes for origins of religious spreads alongside the actual origins of religious spread (see Online Appendix A for an explanation of how each actual origin was selected with reference to the historical literature). We also show the p-values for a test that the estimated and actual coordinates are the same and show that none of the differences are statistically significant.
Source: Authors' calculations.

Both estimates, calibrated using either Buddhism or Islam, are located in south-central Asia, consistent with the hypothesized location of the origin of Hindu scripture (Online Appendix Figure C18). The Islam estimate is slightly farther east than the Buddhism estimate and narrowly leaves out the BMAC, which is one of the hypothesized regions of Hindu scripture, but does fully overlap with the Indus Valley region, which is the other main hypothesis (Online Appendix Figure C18a). However, the estimate calibrated with Buddhism fully overlaps with both regions (Online Appendix Figure C18b). Regardless of the calibration used, the estimates rule out the Pontic Steppe region, which is typically thought of as the origin of Hinduism itself; there is no overlap in either case.[47,48]

Given these patterns, it is not surprising that the distances from our estimates to the history-literature-based estimates are smaller in the case of the origin of scripture compared to the origin of the religion itself. This can be seen in Table 6, Columns (5) and (6). Indeed, the distance to the origin of scripture is 470km if we rely on the Buddhism calibration, and 1,100km if we rely on the Islam calibration. These estimates are larger than for each of the other religions, perhaps reflecting both the greater uncertainty associated with the history literature, and undoubtedly more measurement error associated with loanwords that would have had to have been borrowed so far into the distant past. In any case, despite these distances being larger for Hinduism, they remain much smaller than the comparable distances to the origin of the religion itself. With the Buddhism estimate, the distance to scripture is more than six times closer, and for Islam, it remains more than two and a half times closer. As before, in both cases, the difference is significant well beyond the 1 percent level.

Our conclusion, therefore, is consistent across Christianity, Judaism, and Hinduism. In each case, we find that the language-based estimates are significantly closer to the origin of the religion's scripture than to the origins of the society in which the religion started. While this nuance may help to explain some of the discrepancies that have caused disagreements

[47] We present a series of robustness checks in the Online Appendix Figure C19 as well. In Figure C19a, we show that the estimated coordinate is in essentially the same place when we calibrate using a linear specification. In Figure C19b, we show robustness to a quadratic specification. In Figure C19c, we calibrate using a linear dependent variable instead of the log-dependent variable. In Figure C19d, we calibrate using Betweenness Centrality instead of Eigenvector Centrality, while in Figure C19e, we examine Degree Centrality instead of Eigenvector Centrality. In all cases, the estimated origin location is essentially unchanged.

[48] In Online Appendix Figure C20, we also demonstrate robustness to various alternate clustering algorithms. Again, the estimated origin location is essentially unchanged.

in the history literature, it also stands in stark contrast to early proponents of using etymology to trace historical phenomena, who argued explicitly that linguistic analyses "serve best for determining the origin of peoples" (Leibniz 1996 translation, p. 285).

CONCLUSION

This article empirically assesses the validity of using language etymology to make inferences about the origins of historical phenomena and provides some suggestive evidence that the methodology serves better to identify *spread* rather than the *origin* of the phenomenon itself. To do this, we implement two empirical tests, applied to the historical origins of religion. The first is to test, in the case of Islam and Buddhism, which have straightforward and uncontested origins, whether a fully automated analysis can locate the latitude and longitude of the origins of these religions in the correct places. The second is to test, in cases where the origin of the religion differs from the origin of scripture, whether etymology-based estimates are closer to the former than the latter.

We can, with reasonable accuracy, estimate the origins of both Islam and Buddhism using only information on how words sound and what they mean. In doing so, we present the first quantitative evidence that linguistic analysis can be used in an empirically rigorous way to reconstruct history. Since our approach is entirely empirical—from the identification of religious words to the estimation of their etymology and their link with geographic coordinates—we avoid the main critique associated with using language to reconstruct history. Namely, that it is too open to interpretation by researchers. Furthermore, the estimates for each of Judaism, Christianity, and Hinduism suggest that, at least in the case of religion, language captures the origin of a standardized body of thought more accurately than sacred figures or religious origins. This stands in contrast to the traditional argument in favor of etymology-based historical reconstruction.

While the article focuses on religion, the ability to reconstruct history—at scale—in the absence of detailed primary sources may make the study of questions and contexts that were previously impossible to explore more feasible. That said, there may be important contextual details that are important for the success of the methodology, causing the estimates to be particularly accurate in the case of religion. While we leave the generalizability of the methodology to future work, our approach may be applicable to other questions in economic history when identifying the

origin of the spread of an idea or innovation is of interest. Our approach uses a single measure of linguistic transmission and is therefore most applicable to the location of origin rather than the time of origin.[49] One potential example would be to understand whether slavery or other social institutions had origins within colonized regions or whether they were colonial imports. To the extent that this can be applied more generally, it could help illuminate the histories of peoples, places, and phenomena for which records have been ignored or destroyed.

REFERENCES

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, and Santiago Pérez. "Automated Linking of Historical Data." *Journal of Economic Literature* 59, no. 3 (2021): 865–918.

Ager, Simon. "Omniglot." International Phonetic Alphabet (IPA), 2019. Available at https://www.omniglot.com/writing/ipa.htm.

Alesina, Alberto, and Paola Giuliano. "Culture and Institutions." *Journal of Economic Literature* 53, no. 4 (2015): 898–944.

Algeo, John. *Fifty Years among the New Words: A Dictionary of Neologisms, 1941–1991*. Centennial series of the American Dialect Society. Cambridge: Cambridge University Press, 1993.

Armstrong, Karen. *Islam: A Short History*. New York: Modern Library, 2001.

Assael, Yannis, Thea Sommerschield, Brendan Shillingford, Mahyar Boardbar, et al. "Restoring and Attributing Ancient Texts Using Deep Neural Networks." *Nature* 603, no. 7900 (2022): 280–83.

Bailey, Martha J., Connor Cole, Morgan Henderson, and Catherine Massey. "How Well Do Automated Linking Methods Perform? Lessons from US Historical Data." *Journal of Economic Literature* 58, no. 4 (2020): 997–1044.

Baledent, Anaëlle, Nicolas Hiebel, and Gaël Lejeune. "Dating Ancient Texts: An Approach for Noisy French Documents." In *Proceedings of LT4HALA 2020 – 1st Workshop on Language Technologies for Historical and Ancient Languages*, edited by Rachele Sprugnoli and Marco Passarotti, 17–21. Marseille, France: European Language Resources Association (ELRA), 2020. Available at https://aclanthology.org/2020.lt4hala-1.3.

Barjamovic, Gojko, Thomas Chaney, Kerem Coşar, and Ali Hortaçşu. "Trade, Merchants, and the Lost Cities of the Bronze Age." *Quarterly Journal of Economics* 134, no. 3 (2019): 1455–503.

Becker, Sascha O., and Luigi Pascali. "Religion, Division of Labor, and Conflict: Anti-Semitism in Germany over 600 Years." *American Economic Review* 109, no. 5 (2019): 1764–804.

[49] This would require a list of seed words relevant to the topic in question, like our list of religious seed words. This would also require a known origin that follows a process of spread similar to that of the unknown origin. Since the approach is based on linguistic transmission, it is particularly suited to applications where the topic involves new words, rather than re-interpretations of existing words.

Becker, Sascha O., and Steven Pfaff. "Church and State in Historical Political Economy." In *The Oxford Handbook of Historical Political Economy*, edited by Jeffery Jenkins and Jared Rubin, 925–44. Oxford: Oxford Academic, 2022.

Ben Hamed, Mahe. "Phylo-Linguistics: Enacting Darwin's Linguistic Image." In *Handbook of Evolutionary Thinking in the Sciences*, edited by Thomas Heams, Philippe Huneman, Guillaume Lecointre, and Marc Silberstein. Dordrecht, Netherlands: Springer Nature, 2014.

Bloomfield, Leonard. "Linguistic Aspects of Science." *International Encyclopedia of Unified Science* 1, no. 4 (1939): viii–59.

Blouin, Arthur, and Julian Dyer. "How Cultures Converge: An Empirical Investigation of Trade and Linguistic Exchange." Working Paper, University of Toronto, 2021.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. "Enriching Word Vectors with Subword Information." In *Transactions of the Association for Computational Linguistics vol. 5*, edited by Lillian Lee, Mark Johnson, and Kristina Toutanova, 135–46. Cambridge: MIT Press, 2017.

Campo, Juan E. "Encyclopedia of Islam." Infobase Publishing, 2009. Available at https://archive.org/details/encyclopedia-of-islam_202006.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. "SMOTE: Synthetic Minority Over-sampling Technique." *Journal of Artificial Intelligence Research* 16 (2002): 321–57.

Coleman, Robert. "Book Review of Archaeology and Language by Colin Renfrew." In *Current Anthropology* 29 no. 3. Chicago: University of Chicago Press, Wenner-Gren Foundation for Anthropological Research, 1988. Available at http://www.jstor.org/stable/2743460.

Diebold, R. E. "Linguistic Paleontology." In *The Encyclopedia of Language and Linguistics*. Oxford and New York: Pergamon Press, 1994.

Dyer, Julian, and Arthur Blouin. "[Replication package for] Reconstructing History: Using Language to Estimate Religious Spread." Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2025-06-17. https://doi.org/10.3886/E233003V1

Feigenbaum, James J. "Automated Census Record Linking: A Machine Learning Approach." Boston: Boston University Libraries, OpenBU, 2016. Available at https://hdl.handle.net/2144/27526.

Frankopan, Peter. *The Silk Roads: A New History of the World*. New York: Knopf Doubleday Publishing Group, 2016.

Giorcelli, Michela, Nicola Lacetera, and Astrid Marinoni. "How Does Scientific Progress Affect Cultural Changes? A Digital Text Analysis." *Journal of Economic Growth* 27, no. 3 (2022): 415–52.

Haspelmath, Martin, and Uri Tadmor, eds. *Loanwords in the World's Languages: A Comparative Handbook*. Berlin, Germany: De Gruyter Mouton, 2009.

Lehmann, W. P. "The System of Sonants and Ablaut in Kartvelian Languages: A Typology of Common Kartvelian Structure." *Language* 4, no. 2 part 1 (1968): 404–07.

Leibniz, Gottfried Wilhelm. *New Essays on Human Understanding*. Cambridge: Cambridge University Press, 1996 (translation).

Lemaitre, Guillaume, Fernando Nogueira, and Christos K. Aridas. "Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning." *Journal of Machine Learning Research* 18, no. 17 (2017): 1–5.

Lewis, Paul M. *Ethnologue*: *Languages of the World,* 16th ed. Dallas: SIL International, 2009.

List, Johann-Mattis, Shijulal Nelson-Sathi, Hans Geisler, and William Martin. "Networks of Lexical Borrowing and Lateral Gene Transfer in Language and Genome Evolution." *BioEssays* 36, no. 2 (2014): 141–50.

Lowes, Sara, and Eduardo Montero. "The Legacy of Colonial Medicine in Central Africa." *American Economic Review* 111, no. 4 (2021): 1284–314.

Mackintosh-Smith, Tim. *Arabs: A 3,000-Year History of Peoples, Tribes and Empires*. New Haven, CT: Yale University Press, 2019.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig "Linguistic Regularities in Continuous Space Word Representations." In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, edited by Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff, 746–51. Atlanta, GA: Association for Computational Linguistics, 2013. Available at https://aclanthology.org/N13-1090.

Mortensen, David R., Siddharth Dalmia, and Patrick Littell. "Epitran: Precision G2P for Many Languages." In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018),* edited by Nicoletta Calzolari et al: Paris, France: European Language Resources Association (ELRA), 2018. Available at https://aclanthology.org/L18-1429/.

Nunn, Nathan, and Leonard Wantchekon. "The Slave Trade and the Origins of Mistrust in Africa." *American Economic Review* 101, no. 7 (2011): 3221–52.

Pascali, Luigi. "Banks and Development: Jewish Communities in the Italian Renaissance and Current Economic Performance." *Review of Economics and Statistics* 98, no. 1 (2016): 140–58.

Price, Joseph, Kasey Buckles, Jacob Van Leeuwen, and Isaac Riley. "Combining Family History and Machine Learning to Link Historical Records: The Census Tree Data Set." *Explorations in Economic History* 80, no. 3 (2021): 101391.

Ross, Gavin J. S. "A Program for Fitting Nonlinear Models, MLP." In *Nonlinear Estimation*. Springer Series in Statistics, ch. 7. New York: Springer, 1990.

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.

Schadeberg, Thilo C. "Loanwords in Swahili." In *Loanwords in the World's Languages: A Comparative Handbook*, edited by Martin Haspelmath and Uri Tadmor, 76–102. Berlin, Germany: De Gruyter Mouton, 2009.

Schmidt, J. Die Verwantschaftsverhältnisse der Indogermanischen Sprachen [The Relationship of the Indo-European Languages]. Leipzig: Hermann Böhlau, 1872.

Swadesh, Morris. "Salish Internal Relationships." *International Journal of American Lingustics* 16, no. 4 (1950): 157–67.

Valencia Caicedo, Felipe. "The Mission: Human Capital Transmission, Economic Persistence, and Culture in South America." *Quarterly Journal of Economics* 134, no. 1 (2019): 507–56.

Valencia Caicedo, Felipe, Thomas Dohmen, Andreas Pondorfer. "Religion and Prosociality across the Globe." Working Paper, University of British Columbia, Vancouver, Canada, May 2021.

Vansina, Jan M. *Paths in the Rainforests: Toward a History of Political Tradition in Equatorial Africa*. Madison: University of Wisconsin Press, 1990.

Wichmann, Søren, Eric W. Holman, and Cecil H. Brown, eds. The ASJP Database (version 17), 2016.

Yu, Xuejin, and Wei Huangfu. "A Machine Learning Model for the Dating of Ancient Chinese Texts." In *2019 International Conference on Asian Language Processing (IALP)*. New York: IEEE, 2019. Available at https://doi.org/10.1109/IALP48816.2019.9037653.